ABSTRACT

Title of dissertation: QUANTIFYING THE IMPACT OF REMOTELY SENSED PHOTOSYNTHETICALLY ACTIVE RADIATION RETRIEVALS ON EMPIRICAL CROP MODELS IN THE UNITED STATES

Meredith Guenevere Longshore Brown
Doctor of Philosophy, 2023

Dissertation directed by: Professor Sergii Skakun
Department of Geographical Sciences

Photosynthetically active radiation (PAR), is an essential component for life on Earth and one of the essential climate variables. Due to the differences in biochemistry, cell structure, and photosynthetic pathways, different plant species absorb PAR with varying efficiency and have evolved to thrive in different conditions, such as direct, intense sunlight or indirect, diffuse light conditions. Ground-based measurements allow for direct estimation of PAR; however, those are available in select locations, e.g. through the Surface Radiation Budget (SURFRAD) Network. Remote sensing-based methods, on the other hand, enable spatially explicit estimates of PAR on a regular basis. Current methods and models for satellite-based PAR retrievals require many ancillary atmospheric datasets as well as a large computing infrastructure. PAR, as one of the parameters influencing plant productivity, has not been previously used in the empirical crop yields and as such can lead to better satellite-based yield estimates. Having the advantages of spatially explicit PAR

estimates, spatial and temporal patterns of the PAR can reveal differences in the land uses and the level of crop productivity. Therefore, the overarching goal of my dissertation is to advance the science of satellite-based PAR estimation and agricultural applications. This is done through the use of machine-learning models to reduce data input requirements for PAR estimation from daily Moderate Resolution Imaging Spectroradiometer (MODIS) acquisitions and by incorporating PAR into the empirical crop yield models over the US. In order to obtain satellite-based PAR estimates without the need for ancillary atmospheric data, I developed an empirical approach making use of machine learning methods as an efficient way to capture the non-linear relationship between top of atmosphere radiance and PAR at the surface. I found that the bootstrap aggregated decision tree (Bagged Tree), Gaussian Process Regression (GPR), and Multilayer Perceptron (MLP) yielded the best results with minimal input and training data requirements with an R^2 of 0.77, 0.78, and 0.78 respectively, and a relative RMSE of 22-23%. While these results underperform compared with the look up table (LUT) approach, it does not require the same atmospheric parameters as input, such as atmospheric water vapor, aerosol optical depth, and others that might not be available in near real time or are only available at coarser spatial resolution. I incorporated MODIS-based PAR estimates into empirical corn and soybean yield models over the US. By explicitly adding PAR into the crop yield models, I found a maximum R^2 of 0.81 and 0.80 for corn and soybean, respectively, whereas models that do not include PAR showed a maximum R^2 of 0.60 for corn and soybean. By adding PAR directly into the empirical yield model and demonstrating additional explained variability, I show that my model is in closer

agreement with process-based models than previous empirical models. I found that MODIS- derived coefficient of absorption of PAR (α_{PAR}), which corresponds to the plant canopy chlorophyll content (CCC) and consequently productivity, corresponds to the ground-based α_{PAR} measurements. Specifically, I found that for the US-Ne sites of corn and soybean fields in Eastern Nebraska R^2 was 0.97 and RMSE was 1.34 (11%) when comparing MODIS-derived α_{PAR} with the in situ measurements. I also found that the relationships between MODIS-based α_{PAR} and CCC for corn and soybean corresponded to the ones obtained from in situ data. The relationships between α_{PAR} and CCC for corn and soybean are distinct due to the different photosynthetic pathways of corn (C4) and soybean (C3), differences in cell structure, and chloroplast distribution between the two crops. Crop yield and productivity are also related to CCC, meaning α_{PAR} can be used as a crop specific indicator of yield. Through this research, I have demonstrated the added value of incorporating PAR directly into crop yield models, by improving crop yield estimates over empirical models based on vegetation indices or surface reflectance alone. The research also provides the basis for further work using crop specific measures of the absorption of PAR into the same empirical models at large spatial scales that were previously impractical due to the spatial discrepancies between in situ- and MODIS- derived measurements.

QUANTIFYING THE IMPACT OF REMOTELY SENSED PHOTOSYNTHETICALLY ACTIVE RADIATION RETRIEVALS ON EMPIRICAL CROP MODELS IN THE UNITED STATES

by

Meredith Guenevere Longshore Brown

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Sergii Skakun, Chair
Professor Louis Giglio
Professor George C. Hurtt
Doctor Mark L. Carroll
Professor Joseph Sullivan

Preface

Chapter 2 contains previously published, jointly authored work for which Meredith G. L. Brown was the primary author. Development, processing, analysis, and writing for all papers published, or submitted was led by Meredith G. L. Brown with contributions from the co-authors listed.

Dedication

Моїй родині, велике спасибі. Дуже вас всіх люблю.

To everyone dedicated to open data, open access, and open source for making the doing of science possible in the first place.

To those who have fought, and continue to fight, for the right to participate in science, and in life itself, I stand in solidarity with you and will do what I can to shoulder the burden with you. You belong here. You matter.

And to Jeff Russo for writing the music that I wrote this dissertation to.

Dh'aindeoin có theireadh e

Acknowledgments

First and foremost, I'd like to thank Drs. Sergii Skakun and Shunlin Liang for the opportunities, guidance, encouragement, and support you have both provided. My infinite thanks to Sergii for sticking with me all this time, I would not have been able to finish this dissertation without you. To the rest of my committee and mentors, Drs. Louis Giglio, George Hurtt, Tao He, and Mark Carroll thank you for your advice, feedback, and motivation over the years in shaping and completing this dissertation. Whenever I was feeling like it wasn't possible to keep going, a conversation with one of you helped me realize, "it's just science."

To my colleagues and friends who have helped me with my research and writing, Drs. Yuhan (Doug) Rao, Joanne Hall, Sasha Tyukavina, Catherine Nakalembe, Nicole Motzer, Corinne Carter, Yi Zhang, Zhen Song, Dongdong Wang, Amy Pickens, and Kelly Anderson, thank you for spending your time, wherever we might be, helping me troubleshoot problems, stay focused, edit and proofread, and generally think through the research presented here.

I would also like to acknowledge the help and support from some of the staff members in the department and other faculty across the university for their invaluable assistance with everything from technical problems, to Human Resources issues, to career advice. Jenny Hu, Fernando Ramirez, Jack O'Bannon, Vivre Bell, Liz Smith, Gina Becton, Mary Mitkish, Vicky Berry, Mona Williams, and Drs. Rachel Haber, Sarah Balcom, and Susan Martin thank you for stepping in to save the day, pointing me in a new direction, or just being a sympathetic ear.

On a personal note, I'd like to thank Joanne, Nicole, and Kelly again, as well as Dr. Amanda Hoffman-Hall, Katie McGaughey, Alexandra Cockerham, Rachel Moore, Erika Dailey, and Cortney Gustafson for the sisterhood and inspiration all these years, having a strong community with you has saved my sanity many times.

Special thanks to Erika for letting me teach karate at Cornerstone Martial Arts and help coach the Naval Academy Karate Team. Having that creative outlet and being part of the CMA family has meant the world to me. Day in and day out I am inspired by the perseverance, grit, and dedication of the Midshipmen to learn and compete in some of the toughest arenas, and getting to work with them has helped me up my own game.

Thanks also to Drs. Amanda Woodward and Jeffrey Yeung. Working with you, Susan, and Doug on the PhD Career Navigator was one of the best, and definitely the most fun, experiences of the PhD. I am so grateful to have met you and to have been able to learn from you, that experience has been so important in my overall career development, it's a time I'll cherish forever.

Last, but not least, I owe my deepest thanks to my family and friends-like-family, without you all this dissertation would never have manifested. Thank you to my parents, Lani Longshore and Stewart Brown, who have always led by example, provided every kind of support, and been my biggest cheerleaders. You have instilled in me the values of persistence, empathy, and innovation, and, whether by nature or nurture, made me stubborn, strong, and independent. Without these qualities, I would have quit before I even got started. To my brother, Alexander Brown, thank you for always being sympathetic and uplifting, for playing silly games with me over

months long text messaging, and for being the best kind of competition and support throughout our lives, you've helped me become a better person every day, ever since childhood.

Thank you to all of my east coast family, especially my grandmother and uncle, Lily and Glenn Longshore, who have always provided a refuge out here when I needed one, and to my aunt and uncle, Ilona and Jeff Thurston, who have treated me like a daughter. I am eternally grateful to you all for how you've taken care of me during my PhD. Thank you to my aunts, uncles, and cousins (who are too numerous to list) who started Brown family game night during COVID-19. This has been the most difficult time to deal with the monomania of finishing a dissertation, and the laughs we have when we're all together (even when we're really miles apart) have made the long, anxiety ridden days much more bearable.

My eternal thanks also go to Joanne, Amanda, Kelly, Doug, Catherine, Erika, Lauren Winning, Ana Echavarri-Dailey, and their entire families for having been like family to me. You all made Maryland my home, and I couldn't have finished this dissertation without all that you've done for me.

Despite the length of this list, it is impossible to acknowledge everyone who has contributed to the completion of this dissertation, and I apologize to those I haven't mentioned by name, from the bottom of my heart I thank you all.

Table of Contents

List of Tables

List of Figures

Chapter 1: Photosynthetically active radiation (PAR) and crop yield modeling using satellite remote sensing

1.1 Motivation and Background

The United States produces over a third of the world's corn and soybean (Wang et al., 2020b; Bagnall et al., 2021), grown primarily in the Midwestern United States and the Ohio River Valley. The US Department of Agriculture National Agricultural Statistics Service (USDA NASS) (USDA NASS, 2023) provides consistent and comprehensive agricultural information for all counties in the United States going back to 1850. In addition to the semi-decadal Census of Agriculture reports, researchers can access the annual survey data through the QuickStats tool (`https://quickstats.nass.usda.gov/`). Total acres of corn (maize) planted in the US rose from 79,551,000 in 2000 to 88,579,000 in 2022, while soybean acres planted rose from 74,266,000 in 2000 to 87,450,000 in 2022. Harvest areas from the most recent Census of Agriculture report are shown in Figures 1.1 and 1.2. Corn and soybean yields over the last two decades represent an economic value of $18.6 billion for corn in 2000 to $91.7 billion in 2022 (not accounting for inflation), and $13.1 billion in 2000 for soybeans to $61.1 billion in 2022. Overall yields increased

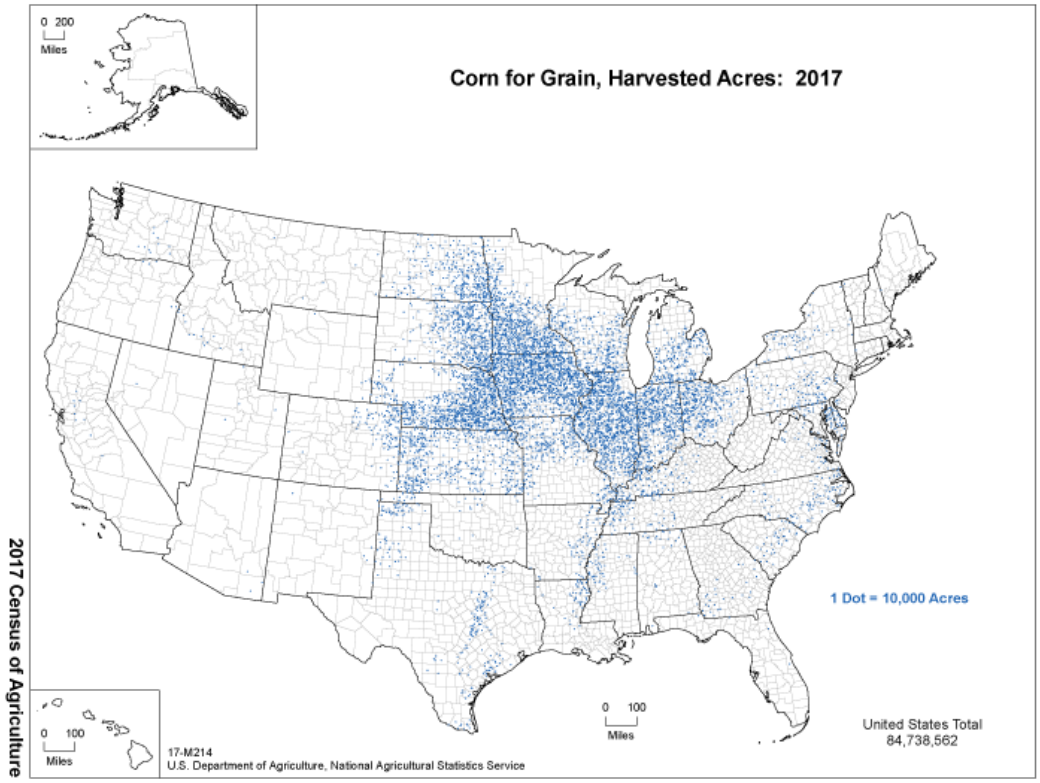

Figure 1.1: Corn for Grain Harvested Acres from the 2017 USDA NASS Census of Agriculture USDA NASS (2023)

by approximately 27% (8.2 t/ha to 10.4 t/ha) for corn and 30% (2.7 t/ha to 3.5 t/ha) for soybeans from 2000-2022, and are shown in Figure 1.3.

According to the US Grains Council (U.S. GRAINS COUNCIL, 2023), in 2022 the US exported 62.7 million tonnes of corn to 62 different countries and the top three were Mexico, China, and Japan. Soybean exports amounted to 71.8 million tonnes in 2022, according to the USDA Economic Research Service and Foreign Agricultural Service as reported by the US Soybean Export Council statement (Kerr-Enskat, 2022). Altogether, this means that modeling and monitoring yields for corn and soybean in the United States are important for both US food security and the

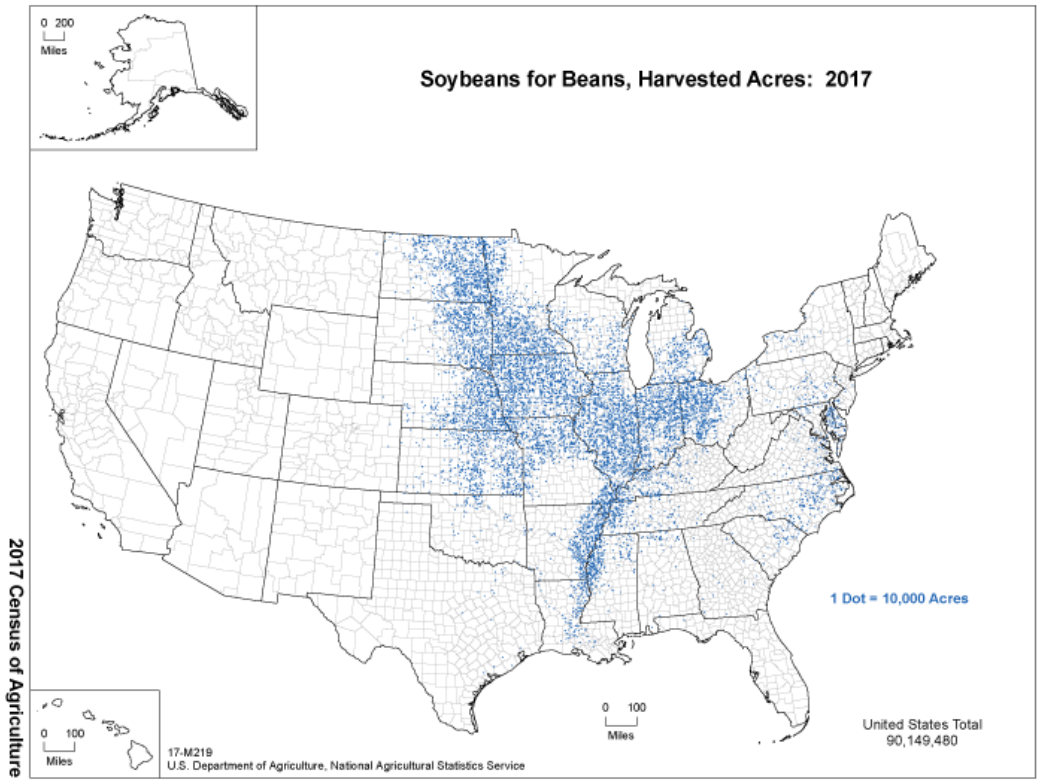

Figure 1.2: Soybeans for Bean Harvested Acres from the 2017 USDA NASS Census of Agriculture USDA NASS (2023)

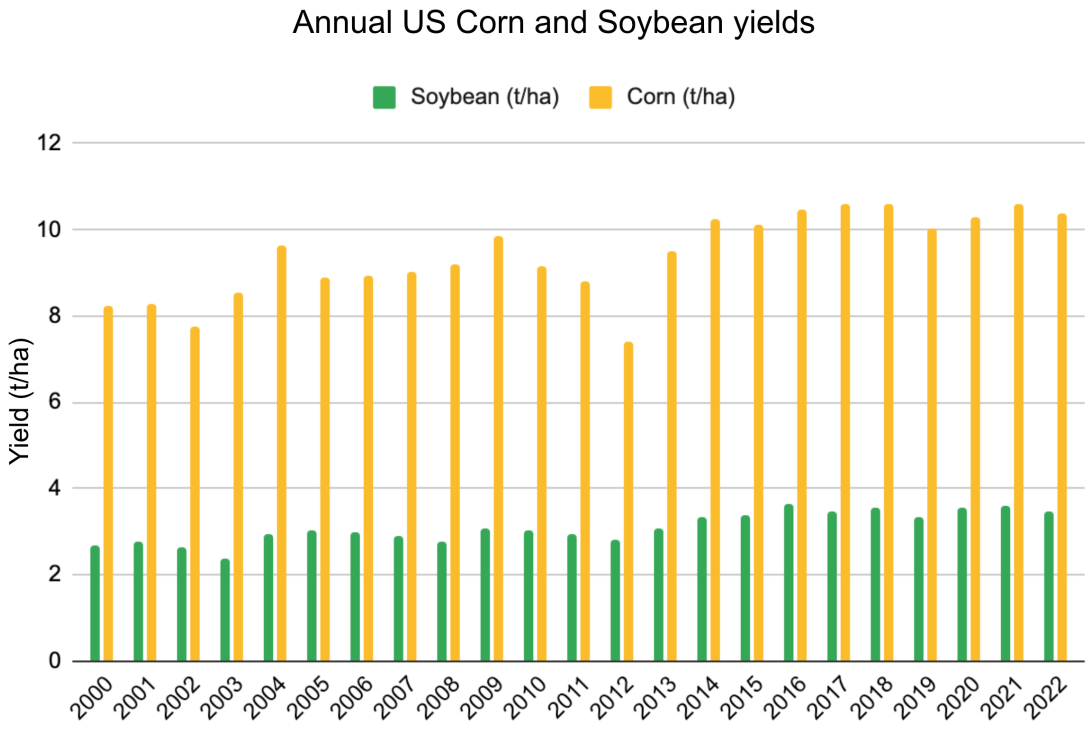

Figure 1.3: 22 year trend of annual US Corn and Soybean yield

economy and for global food markets.

There are numerous studies that model crop yields (Bolton and Friedl, 2013; Basso et al., 2013; Sakamoto et al., 2013; Weiss et al., 2020; Nakalembe et al., 2021). Some methods for crop modeling make use of temperature, precipitation, and certain soil variables (Shirley et al., 2020; Mathieu and Aires, 2018; Park et al., 2005; Lobell et al., 2006) as these physical quantities impact the rate of photosynthesis of different plant species (Medlyn et al., 2002; Mathur et al., 2014), and describe the available water for root systems and the necessary structure, pH, and nutrients required by crops (Munkholm et al., 2013). Other methods use the spectral characteristics of a canopy, e.g., surface reflectance, vegetation indices, and leaf area index (Prasad et al., 2006; Fernandez-Ordoñez and Soria-Ruíz, 2017; Johnson, 2016; Skakun et al.,

2021), as these quantities implicitly contain all the information about the physical conditions of plant or canopy.

Common methods for estimating crop yield from remote sensing data can be divided into physical based models and empirical models. Physical remote sensing based models are developed over specific wavelength domains (e.g., optical, thermal infrared, LIDAR, microwave) and the applicable underlying theory (Weiss et al., 2020). For instance, the Radiative Transfer Model Intercomparison (RAMI) project (Pinty et al., 2001, 2004; Widlowski et al., 2007, 2015) which compares radiative transfer canopy models designed for optical remote sensing observations, such as leaf reflectance and transmittance models (e.g., PROSPECT (Féret et al., 2017, 2021)), plant canopy models such as Scattering by Arbitrarily Inclined Leaves (4SAIL/4SAIL2) (Verhoef and Bach, 2007; Verhoef et al., 2007), the combined PROSPECT and SAIL models, PROSAIL (Jacquemoud et al., 2009; Berger et al., 2018), and soil radiation transfer models, e.g., SOILSPECT (Jacquemoud et al., 1992). These physical models can calculate forward radiative transfers and the radiative transfer inversions, but they are limited by the required input data and perhaps computational capabilities of the user.

Some process-based models can incorporate the climate modeling techniques to capture relationships between crop yield and climate change (Watson et al., 2015). In particular, the Famine Early Warning System Network (FEWSNET) uses coupled land-atmosphere climate models to make near- and long-term predictions of crop conditions. One benefit of using physics, fluid-dynamics, and chemistry-based climate models is that those models can capture temporal scales that data driven

models cannot.

Empirical models, which are often regression models, will use the spectral characteristics of a canopy, e.g., surface reflectance, vegetation indices, and leaf area index (Prasad et al., 2006; Fernandez-Ordoñez and Soria-Ruíz, 2017; Johnson, 2016; Skakun et al., 2021) from remote sensing, as these quantities implicitly contain all the information about the physical conditions of plant or canopy, and calculate (regress) a numerical relationship between remote sensing observations and ground measurements of yields or other biophysical variables. Regression-based methods are data driven, and hence are always limited by the representative nature of available observations. Furthermore, data driven methods, in particular supervised machine learning methods, usually require large amounts of labeled data for training and validation. Such data is not always available for Earth Science applications, especially not for long time periods or with good global representation.

Since the 1970s scientists have been using multispectral satellite data to monitor vegetation (Goward et al., 1985; Justice et al., 1985; Mulla, 2013). Since then the volume of satellite observations and ground measurements has vastly increased, sensors have been developed for rapid deployment in a variety of global locations and applications (Nakalembe et al., 2021). Multispectral remote sensing of vegetation is possible because green vegetation has a very unique spectral curve. The overall reflectance (a measure of the electromagnetic energy that a given surface reflects as a percentage of the amount of energy incident upon it) in the visible range (400 - 700 nm) is quite low due to absorption of light for photosynthesis, while the reflectance in the near infrared range (NIR, 700 - 1300 nm) is quite high, due to

the cell structure of the vegetation. Leaf pigmentation is responsible for variations in reflectance in the visible region, with a small peak typically in the green (500 - 565 nm) region for healthy, active vegetation, which is why most vegetation appears green to the human eye.

Remote sensing of vegetation studies rely heavily on vegetation indices (VIs), which are derived from the unique spectral properties of vegetation as indicators of vegetative health, biomass, and crop yield. In my work, I aim to add PAR as an explicit component to these types of studies. Before doing so, it is important to understand the definitions of different VIs. The simplest VI we can use is a simple ratio between the NIR and visible range, where the visible range is often approximated by the reflectance in the red region ($\sim$625 - 700 nm) for multi-spectral remote sensing.

$$VI = \frac{NIR}{Red}, \tag{1.1}$$

where NIR is the surface reflectance in the near infrared and Red is the surface reflectance in the red region. Simple VI, ranges from 0 to ∞, however, it can be deceptive when overall reflectance is very low or very high. Similarly, we can calculate a simple difference VI (DVI) which ranges from -1 to 1.

$$DVI = NIR - Red \tag{1.2}$$

DVI can also be deceptive when overall reflectance is very high or very low,

therefore, we can normalize DVI as follows:

$$NDVI = \frac{NIR - Red}{NIR + Red}. \tag{1.3}$$

The Normalized Difference Vegetation Index (NDVI) ranges from -1 to 1 and is one of the most commonly used vegetation indices. However, when NIR is very, very high, as can happen in dense, healthy vegetation, NDVI can saturate quickly, thereby creating cases where it is not useful as a crop modeling indicator.

In order to account for the quick saturation of NDVI, a variety of other indices can also be used, such as the Enhanced Vegetation Index (EVI), or the two band version, EVI2, when only reflectance in the red and NIR regions are available.

$$EVI = Gf \frac{NIR - Red}{NIR + C_1 R - C_2 B + L} \tag{1.4}$$

and

$$EVI2 = Gf \frac{NIR - Red}{NIR + C_1 R + L} \tag{1.5}$$

respectively, where B is the surface reflectance in the blue region ($\sim$400 - 485 nm), L is an adjustment factor for the canopy background, C_1 and C_2 are aerosol resistance coefficients, and Gf is a gain factor.

Gross primary production (GPP) is the amount of energy (expressed as biomass) that plant matter creates in a given period. Thus GPP can be used as a proxy for crop yield (Tucker and Sellers, 1986; Reeves et al., 2005; Yuan et al., 2016; Marshall

et al., 2018) when yield data is not available. Before plugging GPP into my yield models, or replacing it with a different proxy for yield, I must understand how GPP is calculated and what exactly it represents.

There are two basic ways to model GPP, either by modeling the biochemical processes that occur in the plant during photosynthesis or by modeling the light-use efficiency (LUE) of an individual plant or the total canopy (Gitelson and Gamon, 2015; Monteith, 1972, 1977). Broadly, LUE is defined by "the ratio of energy output to energy input," (Monteith, 1977) in this case, gross primary production to solar radiation. The original model was developed by Monteith (1972, 1977) and has since been adapted and expanded by others (Xin et al., 2016). According to Monteith (1972) gross primary productivity (GPP) can be expressed by a light-use efficiency (LUE) coefficient times the amount of PAR incident on the canopy, and the fraction of PAR that is absorbed by the canopy ($fPAR$, sometimes written as $FAPAR$ or $FPAR$ in the literature):

$$GPP = LUE \times PAR \times fPAR. \tag{1.6}$$

Most current GPP models are based on light-use-efficiency (Xin et al., 2016; Myneni et al., 2002). With the recent production of the MODIS PAR product (Wang et al., 2020a), PAR can be assimilated into LUE-based yield models at regional and global scales, that previous studies (Johnson, 2016; Skakun et al., 2021), either seasonally or on a daily scale, haven't been able to do. With the addition of PAR to our models, I aim to answer the question, to what degree can PAR improve

agricultural yield models?

Satellite-derived PAR can be useful for modeling crop yield when added to surface reflectance-based crop models (Xin et al., 2016). There is a potential to use this model for near term forecasting of crop yields and to use the model at higher spatial resolutions to improve yield modeling and forecasting for food-insecure regions of the world. However, the differences between the spatial resolution of the PAR product and the spatial resolutions that best support crop yield modeling has yet to be fully determined.

Daily surface reflectance (SR), vegetation indices (VI), leaf area index (LAI), fraction of absorbed PAR (fapar) and/or PAR itself model can be used as indicators for yield correlations (Johnson, 2014, 2016; Gao et al., 2018; Skakun et al., 2021). This can be especially useful in finding the day of the year that is best correlated with crop yield. Studies such as Johnson (2016) used single indicators, such as VI or leaf area index (LAI), here I built a similar model but include multiple input variables, such as surface reflectance in all visible and near-infrared bands, and PAR. The contribution of PAR to vegetation activity and yield has been well studied e.g., (Gitelson et al., 2015; Xin et al., 2016; Alton et al., 2007; Cheng et al., 2015), however up until recently satellite-derived estimates of PAR were not widely available on the global scale or with a decades long time series (Zhao et al., 2013; Wang et al., 2020a).

Vegetation activity (photosynthesis) requires sunlight, precipitation, and favorable temperatures (Nemani et al., 2003; Running et al., 2004; Milesi et al., 2005). The more efficiently PAR is absorbed by cropped vegetation, the higher yields can be (Gitelson et al., 2015; Yuan et al., 2016), which will become increasingly im-

portant as the planet warms and the population rises. Wild (2012) showed that surface shortwave radiation (SSR) trends are associated with increasing trends in both precipitation and near-surface air temperature, while decreasing SSR trends are associated with decreasing precipitation in the northern hemisphere. Other studies have shown how human and natural activity have affected light conditions, particularly with respect to atmospheric aerosols (Roderick et al., 2001; Gu et al., 2003; Rap et al., 2015, 2018), and that with those changing light conditions the amount of carbon removed from the atmosphere during photosynthesis increased (Alton et al., 2007; Mercado et al., 2009; Kanniah et al., 2012; Cheng et al., 2015). However, as the planet changes due to global warming, and the potential for people to attempt various geoengineering strategies (Irvine et al., 2016; Lockley et al., 2020; Liu et al., 2021) to avert some of the adverse effects of climate change, it is increasingly important to be able to monitor and study our cropped vegetation explicitly including radiation as a forcing or indicator.

Surface stations that contain instruments to measure PAR directly can be difficult or expensive to maintain. In cropped fields instruments take up valuable space and it is impractical to have them placed in every field. However, remotely sensed Earth observations are available at a variety of spatial and temporal scales with decades long time series, many with near global coverage. Some remote sensing observations are freely available, while others come from classified or proprietary satellites.

The Moderate Resolution Imaging Spectroradiometer (MODIS) is a prime instrument aboard NASA's Terra and Aqua polar orbiting satellites. MODIS obser-

vations are available four times daily at the global scale, with derived products that range from 250 m to 1 km spatial resolution. The MODIS science team continually works to keep MODIS data and products a reliable source of high quality Earth Observation data for the scientific community. The MODIS data record extends from early 2000 to the present, and while Terra and Aqua have both lasted long past their design lifetimes and their missions will soon be ending, scientists have been working on data fusion with the next generation of comparable NASA observations (Obata et al., 2016; Xiong and Butler, 2020).

MODIS has previously been used in studies for PAR retrievals (Liang et al., 2006; Wang et al., 2020a; Van Laake and Sanchez-Azofeifa, 2004, 2005; Tang et al., 2017), however, these retrievals have primarily been process-based, which are limited by their parameterizations and interpolation schemes. Machine learning methods, on the other hand, in some cases are well suited to resolving non-linearities in interpolation or other limitation from parameterization schemes. Machine learning methods calculate the statistical relationships between the input variables and the desired outputs, or target variables, and should be tested to determine how effectively they can be used for MODIS retrievals of PAR.

As sunlight travels from the top of the atmosphere to the surface, it can be reflected, refracted, absorbed, scattered, or transmitted (Campbell and Wynne, 2011). And thus:

$$I = I_0 e^{-\sigma L}, \tag{1.7}$$

where I is the intensity of the beam at the surface, I_0 is the unattenuated intensity of

the beam at the top of the atmosphere, L is the path length through the atmosphere, and σ is the extinction coefficient, which is equal to the sum of what is scattered or absorbed by the atmosphere (Campbell and Wynne, 2011):

$$\sigma = b_m + b_p + b_n + k, \tag{1.8}$$

where b_m, b_p, and b_n are the coefficients of Rayleigh (scattering off molecules in the atmosphere), Mie (scattering off large particles), or wavelength independent (non-selective) scattering, respectively, and k is the absorption coefficient.

Satellites however, measure radiance at the top of the atmosphere, which includes radiance from the surface as well as the atmosphere, therefore to obtain the amount of radiation at the surface from satellite observations, we must calculate the radiative transfer inversion (Chandrasekhar, 1960) from what is measured by the satellite to what was actually present at the surface.

Incident shortwave radiation can be calculated as follows according to Liang (2005); Liang et al. (2006):

$$F_\lambda(\mu_0) = F_{\lambda,0}(\mu_0) + \frac{r_s\bar{\rho}}{1 - r_s\bar{\rho}}\mu_0 E_0 \gamma(\mu_0), \tag{1.9}$$

where $F(\mu_0)$ is the total incident spectral flux, $F_0(\mu_0)$ is the downward flux without any contribution from the atmosphere, r_s is the surface reflectance, $\bar{\rho}$ is the spherical albedo of the atmosphere, E_0 is irradiance from the sun, $\gamma(\mu_0)$ is the total transmittance through the atmosphere (which could further be partitioned into the

direct and diffuse components), and μ_0 is the cosine of the solar zenith angle. Note, $F_\lambda(\mu_0)$ and $F_{\lambda,0}(\mu_0)$) also have a spectral dependence, λ.

PAR, therefore, can be calculated by integrating all of the incident spectral fluxes over the visible spectrum (400 - 700 nm).

$$PAR(\mu_0) = \int_{400}^{700} F_\lambda(\mu_0) d\lambda. \tag{1.10}$$

Photosynthetically active radiation (PAR) is the portion of sunlight in the visible spectrum, from about 400 - 700 nm that is used for photosynthesis (Anderson, 1971). PAR accounts for approximately 50% of the total radiation received by the surface (Liang et al., 2006).

The coefficient of the absorption of PAR, α_{PAR}, is a dimensionless, semi-analytically modeled measure based on the spectral reflectance of a particular plant species (Gitelson et al., 2019, 2021). Essentially, is a measure of how efficiently and effectively a plant can absorb photosynthetically active radiation for photosynthesis and is defined as follows:

$$\alpha_{PAR} = \frac{\rho_{NIR}}{\rho_{VIS}} - 1, \tag{1.11}$$

where reflectance in the visible spectrum, ρ_{VIS}, is equal to the mean of the reflectance in the red, green, and blue bands:

$$\rho_{VIS} = \frac{\rho_{Red} + \rho_{Green} + \rho_{Blue}}{3}. \tag{1.12}$$

α_{PAR} varies among plant species due to the cell structure, photosynthetic pathway, biochemical properties of the plant, and leaf orientation. It is closely related to both the fraction of absorbed PAR, $fPAR$, and canopy chlorophyll content. α_{PAR} is also likely a stronger indicator of crop yield than vegetation indices for empirical crop models, such as those in Johnson (2016) or (Skakun et al., 2021).

1.2 Dissertation research questions

In this dissertation, I aim to answer the question: To what extent can MODIS-derived PAR be used for studying corn and soybean production in the United States? The research question and design of the dissertation are illustrated in Figure 1.4

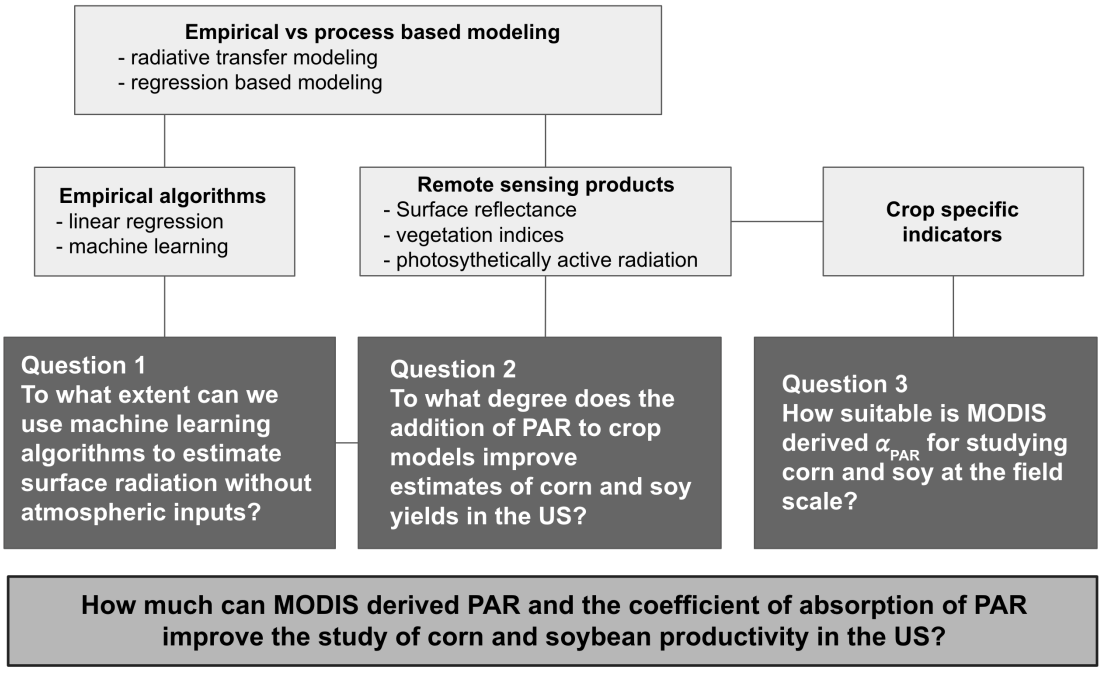

Figure 1.4: Visual description of the dissertation research

In order to answer the broad research question, I break the research down into three components and ask the following three sub-questions. First, to what degree

can an empirical model of surface radiation using limited input data be used to obtain PAR? And how does it compare to physics-based methods? Physics-based methods for retrieving PAR, e.g., Liang et al. (2006), require ancillary atmospheric inputs, from remotely sensed data and reanalysis-based products, such as atmospheric water vapor, aerosol optical depth, and others. These ancillary data are not always available in near real time and they compound the uncertainty of the calculated PAR. I hypothesize that machine learning methods will be able to capture non-linear relationships between top-of-atmosphere radiance measured by satellites and PAR at the surface.

To answer the first question, I chose a selection of machine learning methods to use in my empirical model because advances in computing power and algorithms makes such methods a practical way to capture the nonlinear relationships between TOA reflectance and surface radiation. They are compared to some existing products, including the MODIS product suite of SSR and PAR, MCD18A1 and MCD18A2. The ability to estimate surface radiation without having to rely on all of the necessary ancillary atmospheric data required by radiative transfer would represent a great leap forward in the ability for near real time monitoring or incorporation of PAR into other models.

The second question is, how much yield variability can be explained by adding PAR explicitly to empirical crop yield models of corn and soybean production in the United States at the county scale? Here my hypothesis is that based on the Montieth relationship (Equation 1.6) the incorporation of PAR into the empirical crop yield model will improve crop yield estimates at the county scale in the US. To address this

second question, I take the best available, highest spatial resolution PAR estimates and add them to a county-level crop yield model of corn, and soybean, following the methodology set up by Johnson and Skakun (Johnson, 2016; Skakun et al., 2021). Currently, PAR is only implicitly included in empirical yield models, as PAR affects the greenness of the plant canopy which is seen in the spectral reflectance used by many empirical crop yield models.

And finally, I ask how much variation does MODIS-derived α_{PAR} explain compared to field scale measurements? My final hypothesis is that there will be spatial and temporal variations in the α_{PAR} coefficient due to the different crops and their productivity. Finally addressing the third question, using MODIS surface reflectance and in situ measurements of α_{PAR} and plant chlorophyll content in three test fields with known crop rotations and irrigation methods, I will determine the suitability of using MODIS to calculate α_{PAR} so that it may be incorporated into future empirical crop yield models.

The dissertation is organized as follows: in Chapter 2 the results of an experiment to estimate surface shortwave radiation (SSR) and photosynthetically active radiation (PAR) from top-of-atmosphere (TOA) measurements only using machine learning methods as an alternative to traditional radiative transfer inversion algorithms are presented. Chapter 3 contains the results of a crop modeling study using PAR, surface reflectance, and vegetation indices as indicators of yield. Chapter 4 contains the results of using MODIS-derived α_{PAR} for studying corn and soybean yields at the field- and aggregated field-scales. Finally, overall conclusions, lessons learned, and a vision for future work are presented in Chapter 5.

Chapter 2: Empirical surface radiation retrievals leveraging machine learning methods[1]

2.1 Overview

Satellite-derived estimates of downward surface shortwave radiation (SSR) and photosynthetically active radiation (PAR) are a part of the surface radiation budget, an essential climate variable (ECV) required by climate and vegetation models. Ground measurements are insufficient for generating long-term, global measurements of surface radiation, primarily due to spatial limitations; however, remotely sensed Earth observations offer freely available, multi-day, global coverage of radiance that can be used to derive SSR and PAR estimates. Satellite-derived SSR and PAR estimates are generated by computing the radiative transfer inversion of top-of-atmosphere (TOA) measurements, and require ancillary data on the atmospheric condition. To reduce computational costs, often the radiative transfer calculations are done offline and large look-up tables (LUTs) are generated to derive estimates more quickly. Recently studies have begun exploring the use of machine-learning techniques, such as neural networks, to try to improve computational efficiency.

[1]This work has previously been published as Meredith GL Brown, Sergii Skakun, Tao He, and Shunlin Liang. Intercomparison of machine-learning methods for estimating surface shortwave and photosynthetically active radiation. *Remote Sensing*, 12(3):372, 2020. (Brown et al., 2020)

Here, nine machine-learning methods were tested to model SSR and PAR using minimal input data from the Moderate Resolution Imaging Spectrometer (MODIS) observations at 1 km spatial resolution. The aim was to reduce the input data requirements to create the most robust model possible. The bootstrap aggregated decision tree (Bagged Tree), Gaussian Process Regression (GPR), and Multilayer Perceptron Neural Network (MLP) yielded the best results with minimal training data requirements: an R^2 of 0.77, 0.78, and 0.78 respectively, a bias of 0 ± 6, 0 ± 6, and 0 ± 5 W/m^2, and an RMSE of 140 ± 7, 135 ± 8, and 138 ± 7 W/m^2, respectively, for all-sky condition total surface shortwave radiation and viewing angles less than 55°. Viewing angles above 55° were excluded because the residual analysis showed exponential error growth above 55°. A simple, robust model for estimating SSR and PAR using machine-learning methods is useful for a variety of climate system studies. Future studies may focus on developing high temporal resolution direct and diffuse estimates of SSR and PAR as most current models estimate only total SSR or PAR.

2.2 Introduction

Current satellite-based estimates of surface radiation incorporate atmospheric information in their algorithms, which can be difficult to obtain and propagate error and uncertainty through the algorithm. A popular method for reducing the computational demands of generating a product is to compute the radiative transfer inversions offline and store them in a look-up table (LUT) (Liang et al., 2006; Zhang

et al., 2018; Wang et al., 2020a). LUTs can be generated using in situ data or simulated data, and their major advantage is the ability to do the radiative transfer inversion calculations ahead of time to speed up data generation (Zhang et al., 2018). The major disadvantage is that a LUT must be segmented into bins of a pre-defined size, and then estimates are interpolated between the values in the LUT. The finer the bin segments the larger the LUT, the longer it takes to generate the LUT, and the more time it takes to search the LUT to generate the data of interest. Balancing these requirements is the art of the LUT method. Other methods can also be used to optimize or parameterize a LUT (Zhang et al., 2018), and the computational requirements of these methods is also a limitation of the overall LUT approach.

The aim of this study is to determine if it is reasonable to develop a machine-learning-based model for estimating SSR and PAR from TOA measurements alone. Traditionally surface radiation estimates are generated using physical-based, radiative transfer models (Van Laake and Sanchez-Azofeifa, 2004; Zhang et al., 2015). These models typically require information about the top-of-atmosphere, the atmosphere, and the surface, or they can be parameterized to reduce the ancillary data requirements (Katkovsky et al., 2018). Acquiring this ancillary data introduces sources of potential error, requires heavy-duty computing resources, and is still time intensive (Zhang et al., 2018). Therefore, the goal of this study is to build an empirical model that only requires TOA data as input to reduce these extra sources of potential error, which can be trained and executed quickly and efficiently, while still yielding comparable results to existing methods. I have chosen to test a selection of machine-learning methods (Camps-Valls et al., 2006; Lázaro-Gredilla and Titsias,

2011; Lázaro-Gredilla et al., 2014) in my model to explore how much of the physical processes they can capture as well as possibly improve on computational demands by selecting the smallest reasonable training samples.

For this study, all selected methods are tested with minimal tuning and the best results are identified for further study and development. Here, I use only MODIS TOA measurements and cloud condition, but the model could potentially be adapted to use higher spatial resolution observations such as the Harmonized Landsat Sentinel-2 data (HLS) (Claverie et al., 2018) or they could be adapted for VIIRS (Justice et al., 2013; Skakun et al., 2018) to extend the existing MODIS data record and incorporate further atmospheric or surface information.

2.3 Data

The data sources and years available are shown in Table 2.3. For the first part of the study, the initial intercomparison between machine-learning methods, the surface shortwave radiation (SSR) and photosynthetically active (PAR) models were trained using data from 2005–2009 and independently validated against data from 2010. In the second part of the study, the temporal stability test of the different machine-learning methods in the models, a Leave One Year Out Cross-Validation approach was used, described in Section 2.4.3. All ground truth data are from the Surface Radiation Budget Network (SURFRAD) sites located in the contiguous United States. Each year of data contains approximately 8200 combined satellite overpasses.

Data	Years Avail.	Spatial Res.	Temporal Res.	Citation
MOD/MYD021KM TOA Reflectance	2002-current	1 km at nadir	instantaneous, 1–2-day revisit	(MODIS Science Data Support Team, a,c)
MOD/MYD35 Cloud Mask	2002-current	1 km at nadir	instantaneous, 1–2-day revisit	(Ackerman and Frey, 2015)
MOD/MYD03 Geolocation	2002-current	1 km at nadir	instantaneous, 1–2-day revisit	(MODIS Science Data Support Team, b,d)
SURFRAD	2003-current	10 m footprint	3-min. before 2005, 1-min since 2005	(Augustine et al., 2005)

Table 2.1: Data used for model training and validation.

2.3.1 Remote Sensing

The model inputs are collected from the MODIS top-of-atmosphere (TOA) reflectance from both Terra and Aqua, MOD021KM and MYD021KM respectively, collection 5 (C5), at 1km spatial resolution. I use the reflectance of the first seven bands: red (620–670 nm), near Infrared (841–876 nm), blue (459–479 nm), green (545–565 nm), and the three shortwave infrared bands 1230–1250 nm, 1628–1652 nm, and 2105–2155 nm. Additional inputs to the SSR and PAR models are the satellite viewing geometry: solar zenith angle, satellite view zenith, and the relative angle between the solar and satellite azimuth (relative azimuth angle). I also use the cloud mask (MOD35 and MYD35) as a categorical variable to obtain the cloud condition since no other atmospheric information is explicitly contained in the models.

2.3.2 SURFRAD

The SSR and PAR models are trained and validated using the seven SURFRAD sites in the contiguous United States. The Surface Radiation Budget Network (SURFRAD) consists of seven ground sites in the United States (Augustine et al., 2005) shown in Figure 2.1. The seven SURFRAD sites, which were all installed by 2003, allow for continuous monitoring of direct and diffuse total radiation and PAR at sites in different climate zones, with varying surface types and elevations. The sites have been maintained and updated since their installation, the data is provided in a consistent form with notifications about adjustments and errors to users. While there are other ground sites in the US and other countries as part of other networks, not all of them meet the same standard as the SURFRAD sites, and many were set up as part of short term experiments, and therefore do not have very long data records or the necessary variables available.

The SURFRAD instruments, mounted on platforms 1.5 to 2 m off the ground, and the measurements I used for this experiment are: direct and diffuse solar radiation, and PAR. The direct radiation is measured with a normal incidence pyrheliometer (NIP) which is mounted on a sun tracker, while the diffuse radiation is measured with a shaded pyranometer also attached to the sun tracker. Using the direct and diffuse solar radiation measurements, total SSR is calculated as follows:

$$SSR = R_{dir} * \cos SZ + R_{dif} \tag{2.1}$$

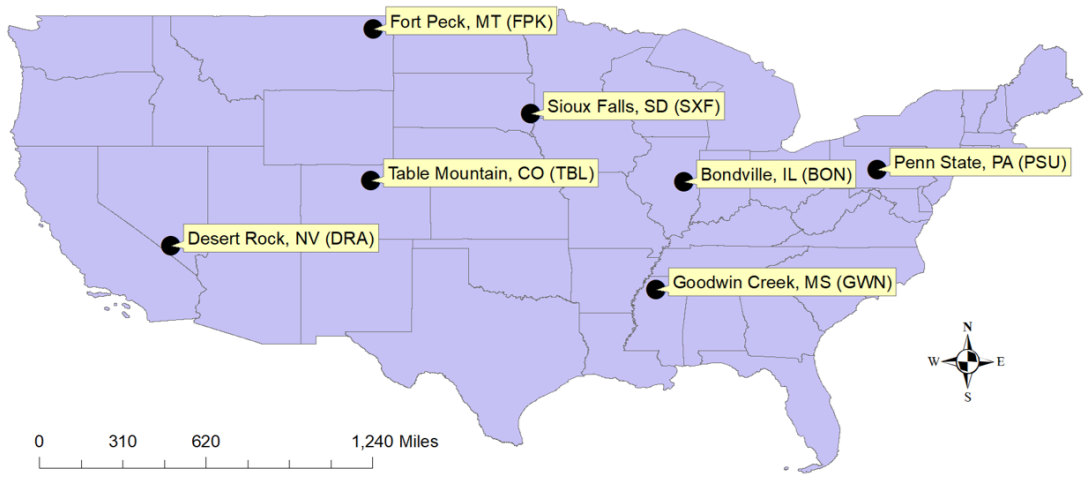

Figure 2.1: Map of the seven SURFRAD sites in the conterminous United States (CONUS).

where R_{dir} is the direct component of radiation, R_{dif} is the diffuse component, and SZ is the solar zenith angle. The uncertainty requirements for the SURFRAD instruments are 2–5% or 15 W/m^2 whichever is larger, to meet the World Climate Research Program specifications (WMO).

2.3.3 Training and Model Validation Data Sets

Prior to training, TOA reflectance from MOD021KM(C5) and MYD021KM(C5) and the overpass times are extracted for pixels containing the location of surface sites. I take a ±15 min temporal average of the SURFRAD data for each satellite overpass. For each site, only one pixel is selected, and no spatial averaging is done at this time following the methods of Zhang et al. (Zhang et al., 2018) and Carter and Liang (Carter and Liang, 2019). Over all sites and all years of data, I have a total of 51,142 data pairs.

2.4 Methods

In this study, four "families" of methods were tested, namely linear methods, decision tree methods, Neural Network-based methods, and kernel-based methods. Below is a brief overview of these types of methods.

2.4.1 Modeling SSR and PAR with Machine-Learning Methods

2.4.1.1 Linear Methods

Regularized Linear Regression (Bishop, 2006) is used as the benchmark method here because it is the simplest, most straightforward method I can use, and one of the most transparent as it gives the most information about the relative importance of the input variables on the model output.

Two additional linear methods were tested, Least Absolute Shrinkage and Selection Operator (LASSO) (Santosa and Symes, 1986; Tibshirani, 1996) and Elastic Net Regularization (ELASTIC NET) (Zou and Hastie, 2005; Hastie et al., 2009). The LASSO method is a type of feature selection and regularization method, which in its simplest form is a type of least squares regression model. ELASTIC NET is a method which includes both the feature selection and regularization of the LASSO method, as well as ridge regression, both methods are supposed to improve prediction accuracy, especially for ill-posed problems.

2.4.1.2 Decision Tree Methods

Decision Trees are a type of non-parametric, supervised learning methods. They approximate a function by incrementally creating a set of if-then-else rules while breaking data sets into increasingly smaller subsets. A Bootstrap Aggregated (Bagged) Decision Tree (Breiman, 1996) is a special case of the ensemble approach applied to the basic decision tree method that can reduce variance and avoid over-fitting. The method works by sampling the original training set for each new tree to create an ensemble of trees from which predictions can be made.

2.4.1.3 Neural Networks

The feed-forward multi-layer perceptron (MLP) is one of the most common neural networks. In this method, the inputs are fed through the hidden layers and connected to the outputs through a series of weights. The outputs of each layer are compared to the desired outputs and fed back through the network, adjusting the weights each time, until the error function has been minimized (Bishop, 2006; Camps-Valls et al., 2006; LeCun et al., 2015; Kussul et al., 2017).

2.4.1.4 Kernel Methods

Of the many different types of kernel methods, Gaussian Process Regression (GPR) (Bishop, 2006; Lázaro-Gredilla and Titsias, 2011) sometimes also known as kriging, is a type of distance weighting machine-learning algorithm that makes use of an assumed Gaussian probability distribution to make its predictions. This

feature of the method requires small training sample sizes lest the model become too cumbersome.

2.4.2 Data Filtering, Parameter Tuning, and Training

My aim is to create an all-sky model, therefore I include all sky conditions identified according to the MODIS cloud mask. The inputs to the model are solar zenith angle (SZ), view zenith angle (VZ), relative azimuth (AZ), reflectance in the first seven top-of-atmosphere (TOA) MODIS bands, and the coded cloud condition described in Section 2.3.3. Training data is filtered so that the training set contains only pixels whose cloud flag matches the expected amount of ground-measured radiation are used for training; however all valid pixels are included in the model validation data, so there may be mismatches between the satellite cloud mask and the ground observed cloud condition.

Data viewed above 55° is discarded. Due to the bowtie effect, the pixel size at such extreme viewing angles is much larger than the pixel size at and nearer to nadir (Campagnolo and Montano, 2014). Furthermore, the additional path length through the atmosphere at such extreme viewing angles contaminates the pixel compared to small viewing angles. The MODIS team recommends against using pixels at such high viewing angles due to data quality issues associated with the extremity of the viewing angle (MODIS Science Data Support Team, a).

Each method has a different optimal training sample size, for example, a Neural Network benefits from the largest training sample available, whereas a Gaussian

Process Regression is optimized for small training sample sizes (1000–2000 points), therefore for each method, the model is allowed to separate its training and internal testing samples randomly, according to its optimal parameters. For this reason, several tests are done leaving specific data out for independent cross-validation and intercomparison.

Each method also has a different set of tunable parameters. The linear methods have at most one or two parameters to tune, whereas the Neural Network has several, including number of hidden layers, number of neurons per layer, number of epochs, and the Kernel Ridge Regression and Gaussian Process Regression have a gamma parameter that defines the kernel. For each combination of tunable parameters, tests were performed to find the final combination that yielded the best results with the smallest computational requirements.

During parameter tuning tests, the final combination of parameters selected had to maximize R^2, while minimizing RMSE and training time. The Bagged Tree reaches its peak performing parameters with only 200 trees in the bag, yet I tested up to 2000 trees to see if I could get any improvement on RMSE. It is possible that with more than 200 bags in the tree, the method overfits, other works suggest that the number of trees in the bag should depend on the number of features in the model (Latinne et al., 2001; Oshiro et al., 2012; Perner, 2012). In my model I have 14 input variables, thus it should be expected that a relatively small number of trees should be optimal.

For the final parameter selection, I used a Multi-layer Perceptron Neural Network with 1 hidden layer, 14 neurons (one for each input), 100 training epochs, and is

trained with the Levenberg-Marquardt backpropagation method. The Bagged Tree and Boosted Tree methods use 200 trees/bag. The Kernel Ridge Regression and Gaussian Process Regression use a radial basis (RBF) kernel. Other kernel functions tested were not successful.

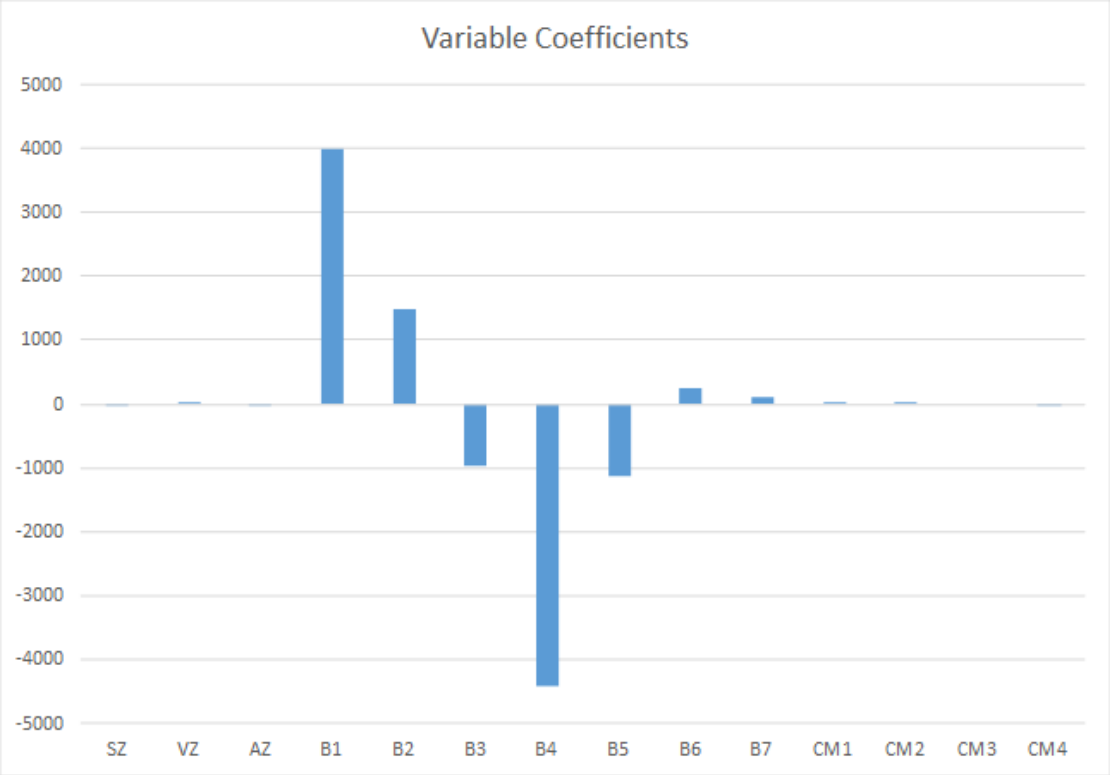

Figure 2.2: Relative importance of the model input variables.

The RLR method coefficients, shown in Figure 2.2, show that the MODIS red band (620–670 nm), where the peak solar energy is measured, and green band (545–565 nm), are the most influential input variables, and further that as reflectance in the blue (459–479 nm), green, and SWIR (1230–1250 nm) bands increases, the estimated SSR or PAR decreases, showing that these bands are the most sensitive to aerosols in the atmosphere that might lead to the global dimming phenomenon (Wild, 2012).

2.4.3 Model Cross Validation

Each machine-learning method has an optimal training sample size. Linear regression and neural networks perform best with large training samples, whereas Kernel Ridge Regression (KRR) and Gaussian Process Regression (GPR) are better suited for smaller training data sets. In order to make the best use of each method I use different training samples sizes according to the method and then use three different methods for model validation and intercomparison.

Model Validation Method (1): Data from 2005–2009 are used for the training set, while data from 2010 is used for independent validation and intercomparison. Using this training and model validation method, I have 42,754 data pairs available for training and 8388 for validation.

Model Validation Method (2): I use the Leave One Year Out Cross-Validation (LOYOCV) method, a type of k-fold cross-validation, where I hold one year out and repeat the training and cross-validation 6 times, each time using five years for training and holding one year out for model validation. On average, there are 1217 data pairs per site per year, meaning each iteration has an average of 42,618 data pairs available for training and 8523 for validation.

Model Validation Method (3): I use the Leave One Station Out Cross-Validation (LOSOCV) method, a similar type of k-fold cross-validation, and train on six of the seven SURFRAD stations and hold one out for cross-validation, iterating through just as I did for the LOYOCV. On average, each iteration in this model validation method has 43,836 data pairs available for training and 7306 for validation. By us-

ing this type of cross-validation, I build an ensemble from which I can determine the spatial and temporal stability. In order to evaluate the different machine-learning methods and compare them to each other I calculate R^2_{adj}, RMSE, and bias as follows:

$$R^2 = \frac{\sum \left(R_{est} - \overline{R_{obs}}\right)^2}{\sum \left(R_{obs} - \overline{R_{obs}}\right)^2} \tag{2.2}$$

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{2.3}$$

$$RMSE = \sqrt{\frac{\sum \left(R_{est} - R_{obs}\right)^2}{n}} \tag{2.4}$$

$$Bias = \frac{\sum \left(R_{est} - R_{obs}\right)}{n} \tag{2.5}$$

where R_{est} are the modeled surface radiation (either SSR or PAR), R_{obs} are the ground-measured radiation data, $\overline{R_{obs}}$ is the mean of the ground-measured radiation, n are the number of data pairs, and p is the number of input parameters.

2.5 Results

2.5.1 Model Performance

The results of the nine machine-learning methods, shown in Table 2.5.1 show that the linear methods and a single decision tree do not simulate the ground observed SSR and PAR as well as the non-linear methods. The best methods for SSR

and PAR are the bootstrap aggregated (BAGGED TREE) decision tree (Figures 2.3 and 2.4), the multi-layer Neural Network (MLP) (Figures 2.5 and 2.6), and the Gaussian Process Regression (GPR) methods (Figures 2.7 and 2.8). All methods are in good agreement, but show some deviation from the 1:1 line, especially at low values of PAR or SSR. The GPR method best corrects this effect, but the overall spread of the modeled radiation at low values increases somewhat compared the MLP and BAGGED TREE methods.

Method	**SSR** R^2_{adj}	**SSR RMSE** ($\mathbf{W/m^2}$)	**SSR Bias** ($\mathbf{W/m^2}$)	**PAR** R^2_{adj}	**PAR RMSE** ($\mathbf{W/m^2}$)	**PAR Bias** ($\mathbf{W/m^2}$)
RLR	0.68	170 (29%)	−11	0.70	69 (29%)	−3
LASSO	0.68	170 (29%)	28	0.70	70 (29%)	11
ELASTIC NET	0.69	170 (29%)	29	0.70	70 (29%)	11
DECISION TREE	0.62	190 (31%)	−8	0.60	82 (31%)	−3
BAGGED TREE	0.77	144 (23%)	−8	0.76	61 (23%)	−2
BOOSTED TREE	0.73	155 (25%)	−11	0.73	65 (24%)	−3
MLP	0.78	138 (22%)	−4	0.78	59 (22%)	−1
KRR	0.75	149 (24%)	−7	0.75	62 (23%)	−1
GPR	0.78	140 (23%)	−5	0.78	59 (22%)	−2

Table 2.2: Model Validation Method 1 results for both SSR and PAR.

2.5.2 Time Series and Site Analysis

During training, two types of cross-validation were conducted to test the robustness of the methods. First, in Model Validation Method 2 of my analysis, I test the temporal stability of the model methods. In Leave One Year Out Cross-Validation (LOYOCV), I iteratively train the model on only five of the six years and use the last year held out for cross-validation. In this way, I test the temporal robustness of the model methods, the statistics are given in Tables 2.5.2 and 2.5.2,

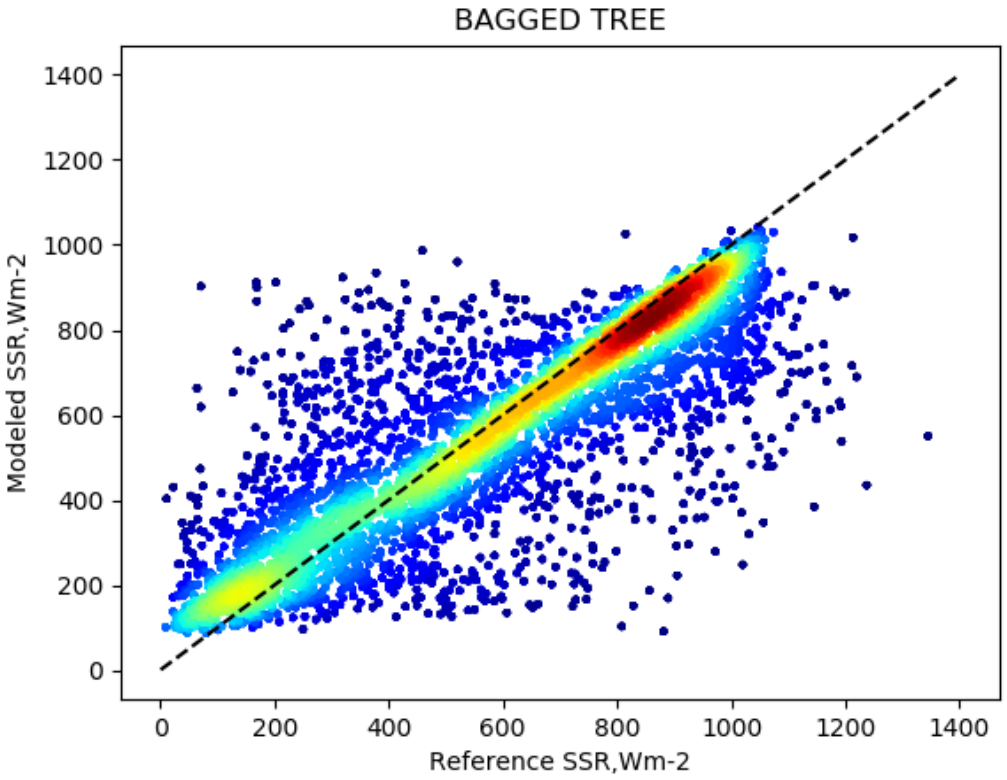

Figure 2.3: Model Validation Method 1 results for SSR BAGGED TREE, $R^2_{adj} = 0.77$, RMSE = 144 (W/m^2)(23%)

and they show that the methods are temporally stable.

For the LOYOCV, I find that for these six years of data, the model is temporally stable, and there are no outlier years. I find comparable results for the PAR LOYOCV. Keeping in mind that since PAR is approximately half of SSR, the RMSE and bias for the PAR are relatively the same as for SSR. Second, in Model Validation Method 3, I tested the spatial stability, using the Leave One Station Out (LOSOCV) cross-validation approach, similar to Model Validation Method 2, I iteratively train on only six of the seven SURFRAD sites and use the site held out for cross-validation. The statistics are given in Tables 2.5.2 and 2.5.2 and discussed

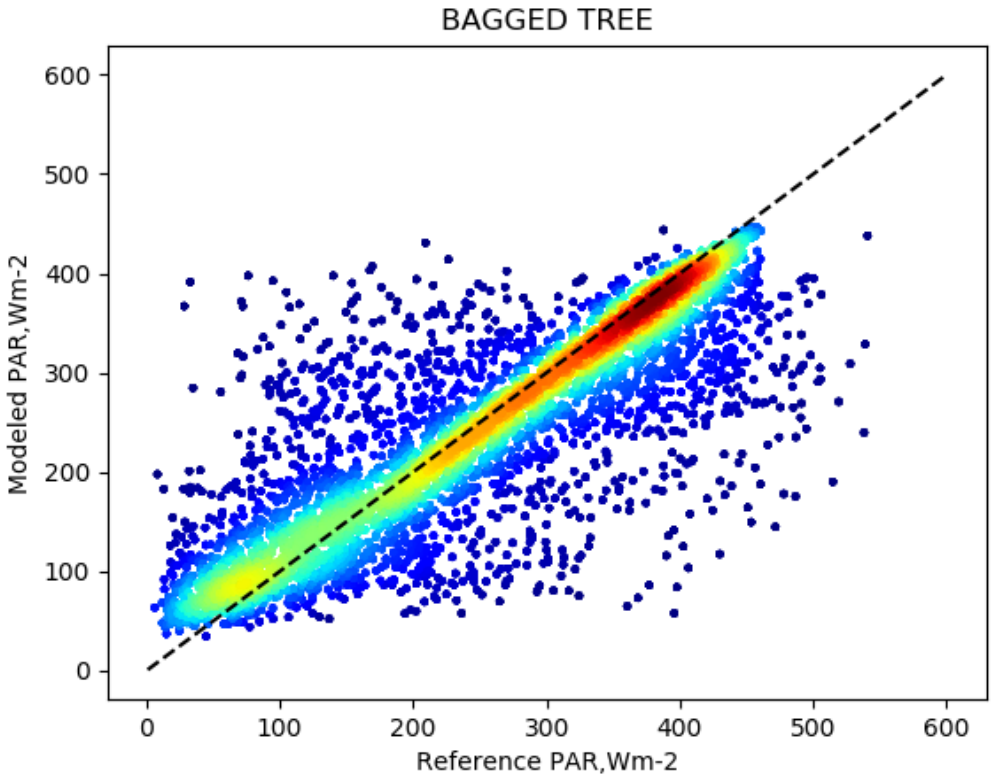

Figure 2.4: Model Validation Method 1 results for PAR BAGGED TREE, $R^2_{adj} = 0.76$, RMSE = 61 (W/m^2) (23%)

in the following sections.

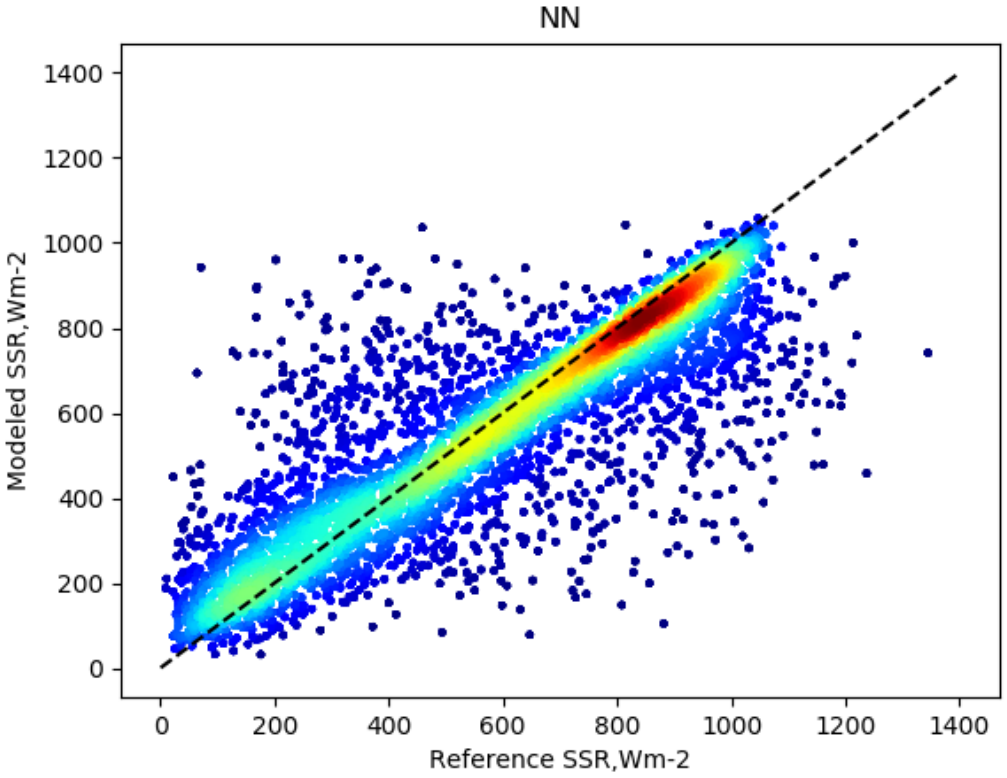

Figure 2.5: Model Validation Method 1 results for SSR MLP, $R^2_{adj} = 0.78$, RMSE = 138 (W/m^2) (22%)

For the PAR LOSOCV, the results are comparable to the SSR, as shown in Table 2.5.2.

2.6 Discussion

The most accurate of the machine-learning methods were the bootstrap aggregated decision tree, the Multi-layer Perceptron Neural Network, and the Gaussian Process Regression. I find that regardless of the chosen method, the model is quite stable when I performed an iterative training and cross-validation through time and space. Among the SURFRAD sites, the Table Mountain site near Boulder, CO

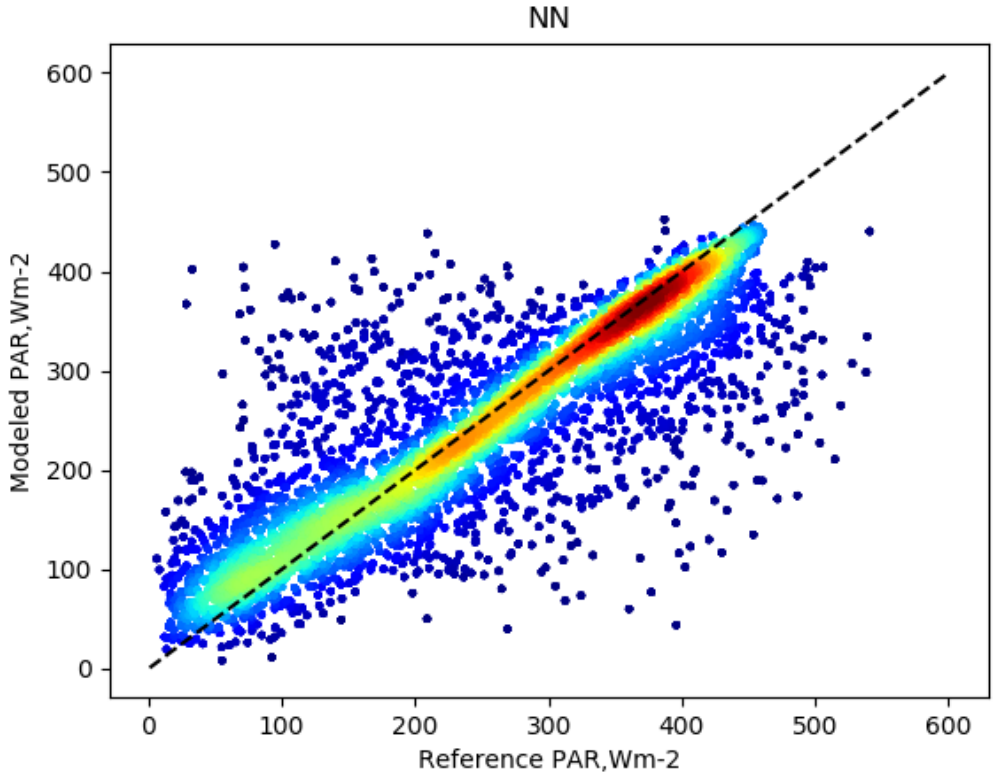

Figure 2.6: Model Validation Method 1 results for PAR MLP, $R^2_{adj} = 0.78$, RMSE = 59 (W/m^2) (22%)

shows considerably different model validation statistics from the other sites that skews the spatial cross-validation somewhat. Including more sites in the training and/or cross-validation may resolve this issue; however ground measurements from other networks in the United States do not have the same data quality or data record as the SURFRAD sites that were designed for long-term radiation monitoring (Augustine et al., 2005).

The optical depth of the atmosphere, whether due to clouds or aerosols, still presents a challenge to this work. The thickness of clouds and aerosols is a major factor in how much radiation can reach the surface (Xu et al., 2011; Lefèvre et al.,

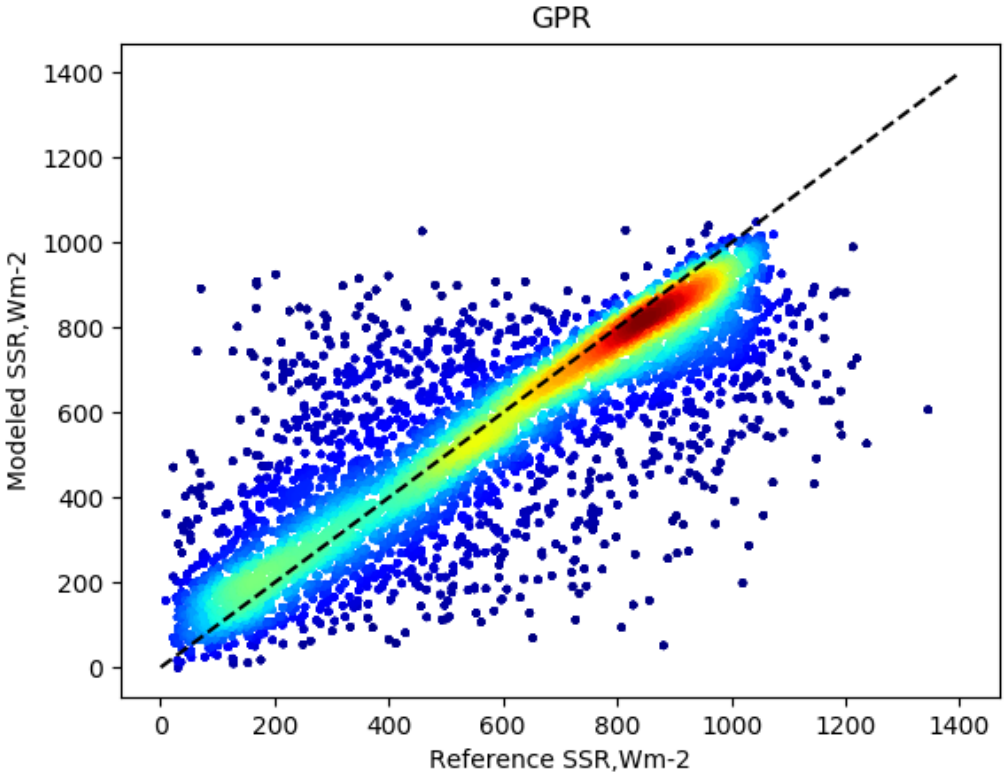

Figure 2.7: Model Validation Method 1 results for SSR GPR, R^2_{adj} = 0.78, RMSE = 140 (W/m^2) (23%)

2013; Xu et al., 2016). The aim of this work was to test if standard machine-learning methods could accurately estimate SSR and PAR without this *a priori* information, and I have shown that they can within 20% error. However, while machine learning can infer statistical relationships and make estimations based on those relationships, the missing information from the model will likely be seen in the comparison between these methods and other satellite estimates based on physical models.

I have reported my results as instantaneous estimates of SSR and PAR, while many other studies report 3-hourly estimates (Zhang et al., 2014; Gui et al., 2010; Zhang et al., 2018). Zhang et al. (Zhang et al., 2018) report RMSE of 12% for their

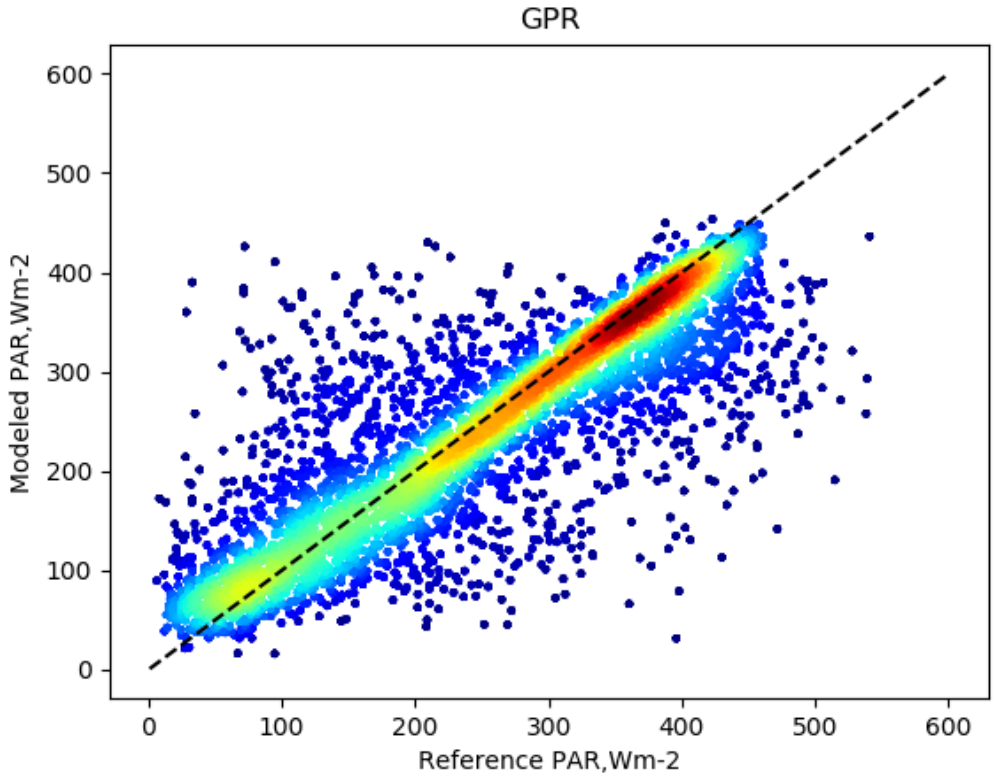

Figure 2.8: Model Validation Method 1 results for PAR GPR, $R^2_{adj} = 0.78$, RMSE = 59 (W/m^2) (22%)

Method	R^2	std	RMSE (W/m^2)	std (W/m^2)	Bias (W/m^2)	std (W/m^2)
RLR	0.62	0.08	183 (30%)	25	5	9
LASSO	0.65	0.07	182 (30%)	18	51	12
ELASTIC NET	0.65	0.07	182 (30%)	19	50	10
DECISION TREE	0.60	0.01	193 (32%)	4	−2	7
BAGGED TREE	0.77	0.01	140 (23%)	6	0	6
BOOSTED TREE	0.73	0.02	151 (25%)	7	1	7
MLP	0.78	0.02	136 (22%)	7	1	6
KRR	0.76	0.02	141 (23%)	7	0	5
GPR	0.78	0.02	138 (23%)	7	0	5

Table 2.3: Model Validation Method 2, Leave One Year Out Cross-Validation (LOY-OCV) results for SSR.

Method	R^2	std	RMSE (W/m^2)	std (W/m^2)	Bias (W/m^2)	std (W/m^2)
RLR	0.63	0.08	78 (30%)	10	2	4
LASSO	0.65	0.06	77 (29%)	8	22	5
ELASTIC NET	0.65	0.06	78 (29%)	8	22	5
DECISION TREE	0.61	0.01	81 (31%)	1	−1	2
BAGGED TREE	0.77	0.02	60 (23%)	2	0	2
BOOSTED TREE	0.73	0.02	64 (24%)	3	0	2
MLP	0.79	0.02	58 (22%)	3	0	2
KRR	0.77	0.02	60 (23%)	3	0	2
GPR	0.77	0.03	59 (22%)	4	-1	2

Table 2.4: Model Validation Method 2, Leave One Year Out Cross-Validation (LOY-OCV) results for PAR.

3-hourly estimates at the SURFRAD sites, while other estimates range from 14–24% at the same sites. The best comparison I can make is to the instantaneous SSR and PAR estimates from the new MODIS suite of products, MCD18. Wang et al. (Wang et al., 2020a) report RMSE between 10–18% at the different SURFRAD sites.

2.7 Conclusions

In this work I tested nine machine-learning methods to model SSR and PAR using minimal input data from the MODIS instrument at 1 km spatial resolution in

Method	R^2	std	RMSE (W/m^2)	std (W/m^2)	Bias (W/m^2)	std (W/m^2)
RLR	0.60	0.09	182 (31%)	33	7	17
LASSO	0.63	0.07	184 (31%)	31	54	36
ELASTIC NET	0.63	0.07	183 (31%)	31	52	36
DECISION TREE	0.51	0.10	214 (36%)	27	-19	55
BAGGED TREE	0.74	0.04	149 (25%)	13	-11	42
BOOSTED TREE	0.70	0.04	155 (26%)	10	-2	24
MLP	0.76	0.04	139 (23%)	12	-8	30
KRR	0.73	0.04	146 (25%)	15	8	13
GPR	0.75	0.04	141 (24%)	15	-2	15

Table 2.5: Model Validation Method 3, Leave One Site Out Cross-Validation (LOSOCV) results for SSR.

Method	R^2	std	RMSE (W/m^2)	std (W/m^2)	Bias (W/m^2)	std (W/m^2)
RLR	0.60	0.09	78 (31%)	14	3	7
LASSO	0.63	0.07	78 (31%)	12	21	16
ELASTIC NET	0.63	0.07	78 (30%)	12	21	16
DECISION TREE	0.54	0.05	88 (34%)	7	−4	17
BAGGED TREE	0.74	0.04	64 (25%)	5	−5	18
BOOSTED TREE	0.71	0.04	67 (26%)	4	0	16
MLP	0.76	0.04	59 (23%)	5	−1	9
KRR	0.67	0.11	72 (28%)	21	−1	4
GPR	0.75	0.03	61 (24%)	5	0	9

Table 2.6: Model Validation Method 3 LOSOCV results for PAR.

order to explore the ability of machine-learning-based, empirical model to estimate surface shortwave radiation (SSR) and photosynthetically active radiation (PAR) using input data from minimal sources to reduce error propagation and computational time. I found that the bootstrap aggregated decision tree (Bagged Tree), Gaussian Process Regression, and Multi-layer Perceptron Neural Network yield the best results with minimal input and training data requirements. I report an R^2 of 0.77, 0.78, and 0.78 respectively, a bias of 0 ± 6, 0 ± 6, and 0 ± 5 W/m^2, and an RMSE of 140 ± 7, 135 ± 8, and 138 ± 7 W/m^2, respectively, for all-sky condition total surface shortwave radiation and viewing angles less than 55°.

Future studies should focus on several areas: 1) Adding more MODIS bands as inputs to the model. While the first 7 MODIS bands cover a large portion of the electromagnetic spectrum, some of the other bands may be more sensitive to aerosols in the atmosphere that would limit radiation from reaching the surface. 2) Adding more training data to the model. My work was aimed at finding the smallest reasonable training sample and the simplest reasonable model design, but more training samples, may improve the model assuming the differences in measurements and calibrations could be well handled. 3) More aggressive filtering of input and training data. My intention was to include as much data as possible; however, starting from an idealized clear-sky model and building a more complex model to handle cloudy-sky cases could be one strategy to improve the model results. 4) Developing high temporal resolution direct and diffuse estimates of SSR and PAR as most current models estimate only total SSR or PAR.

Chapter 3: Incorporating photosynthetically active radiation (PAR) into crop yield models for corn and soybeans in the US

3.1 Overview

Photosynthetically active radiation (PAR), as one of the parameters influencing plant productivity, is not typically explicitly included in satellite remote-sensing based empirical crop yield models, rather these models tend to be based on vegetation indices (VIs) and other spectral properties observed by satellites which implicitly include information about radiation conditions. However, since the release of the official Moderate Resolution Imagining Spectroradiometer (MODIS) PAR product (MCD18A2) (Wang et al., 2020a), long term, global surface radiation data are now available for incorporation into yield models. Having the advantages of spatially explicit PAR estimates, spatial and temporal patterns of the PAR can reveal differences in the land uses and the level of crop productivity. Here I use multiple indicators and their combinations, including MODIS PAR, VI, and surface reflectance (SR) to model crop yields of corn and soybean at the county level in the US from 2001-2020. I find that the addition of PAR to empirical yield models of corn and soybean in the US does increase the adjusted coefficient of determination (R^2_{adj})

compared to models that rely on VI or SR alone. For VI only based models, I find maximum R^2_{adj} around 0.60 for both corn and soybean, and models that include PAR typically improve maximum R^2_{adj} to around 0.80 for both corn and soybean. My findings indicate the value added by incorporating PAR into empirical crop yield models, even at coarse spatial scale in a region where vegetation is not radiation limited. They also suggest there is value for future studies in estimating surface radiation from high spatial resolution satellite data.

3.2 Introduction

The United States produces over a third of the world's corn and soybean (Wang et al., 2020b; Bagnall et al., 2021), grown primarily in the Midwestern United States and Ohio River Valley. The US Department of Agriculture National Agricultural Statistics Service (USDA NASS) (USDA NASS, 2023) provides consistent and comprehensive agricultural information going back to 1850. The economic value of corn and soybean production in the United States has grown by over $115 billion combined in the last 20 years. Annually, the US exports 10-20% of its supply to dozens of countries worldwide, making modeling and monitoring corn and soybean yield important both domestically and internationally.

Common methods for estimating crop yield from remote sensing data can be divided into physical based models and empirical models. Physical models are developed over specific wavelength domains (e.g., optical, thermal infrared, LIDAR, microwave) and the applicable underlying theory (Weiss et al., 2020). For instance,

the Radiative Transfer Model Intercomparison (RAMI) project (Pinty et al., 2001, 2004; Widlowski et al., 2007, 2015) which compares radiative transfer canopy models designed for optical remote sensing observations, such as leaf reflectance and transmittance models (e.g., PROSPECT (Féret et al., 2017, 2021)), plant canopy models such as Scattering by Arbitrarily Inclined Leaves (4SAIL/4SAIL2) (Verhoef and Bach, 2007; Verhoef et al., 2007), the combined PROSPECT and SAIL models, PROSAIL (Jacquemoud et al., 2009; Berger et al., 2018), and soil radiation transfer models, e.g., SOILSPECT (Jacquemoud et al., 1992). These physical models can calculate forward radiative transfers and the radiative transfer inversions, but they are limited by the required input data and perhaps computational capabilities of the user.

Empirical models, or regression models, will use the spectral characteristics of a canopy, e.g., surface reflectance, vegetation indices, and leaf area index (Prasad et al., 2006; Fernandez-Ordoñez and Soria-Ruíz, 2017; Johnson, 2016; Skakun et al., 2021) from remote sensing, as these quantities implicitly contain all the information about the physical conditions of plant or canopy, and calculate (regress) a numerical relationship between remote sensing observations and ground measurements of yields or other biophysical variables. Regression-based methods are data driven, and hence are always limited by the representative nature of available observations.

In this study, I follow the methodology of Johnson (2016), who tested multiple remotely sensed indicators, such as VI or leaf area index (LAI) to model crop yields. Here I built similar models but include multiple input variables, such as surface reflectance in all visible and near-infrared bands, and PAR. The contribution of

PAR to vegetation activity and yield has been well studied e.g., (Gitelson et al., 2015; Xin et al., 2016; Alton et al., 2007; Cheng et al., 2015), however up until recently satellite-derived estimates of PAR were not readily available on the global scale or with a decades long time series (Zhao et al., 2013; Wang et al., 2020a).

3.3 Study area

For the study, I selected the four MODIS tiles (h10v04, h11v04, h10v05, and h11v05) containing the corn and soy belt in the United States, which covers the Midwest and Ohio River valley. My study area extends westward from the Great Plains to the mountain west from Kansas, the Dakotas, Montana, and into Idaho and eastern Washington, as well south and eastward to the mid-Atlantic, Southeastern, and Gulf states.

There are 1150 counties in the study area with reported corn yields, representing over 80% of the nearly 90 million acres of planted corn in the United States. There are 1094 counties with reported soybean yields. Corn yields are typically highest in the central counties of the study area in Illinois, Iowa, Kansas, southern Minnesota, and along the Mississippi River in Arkansas (Figure 3.1). Similarly, soybean yields are typically highest in Illinois, Iowa, Missouri, Arkansas, and Kansas (Figure 3.2). Annual yields for corn and soybean from 2001-2020 are shown in Figures 3.3 and 3.4. The boxplots show the median yield for each county and the 2nd and 3rd quartiles. The whiskers are defined by 1.5 $\times$ the standard deviation and the black dots are yields that fall outside of that range.

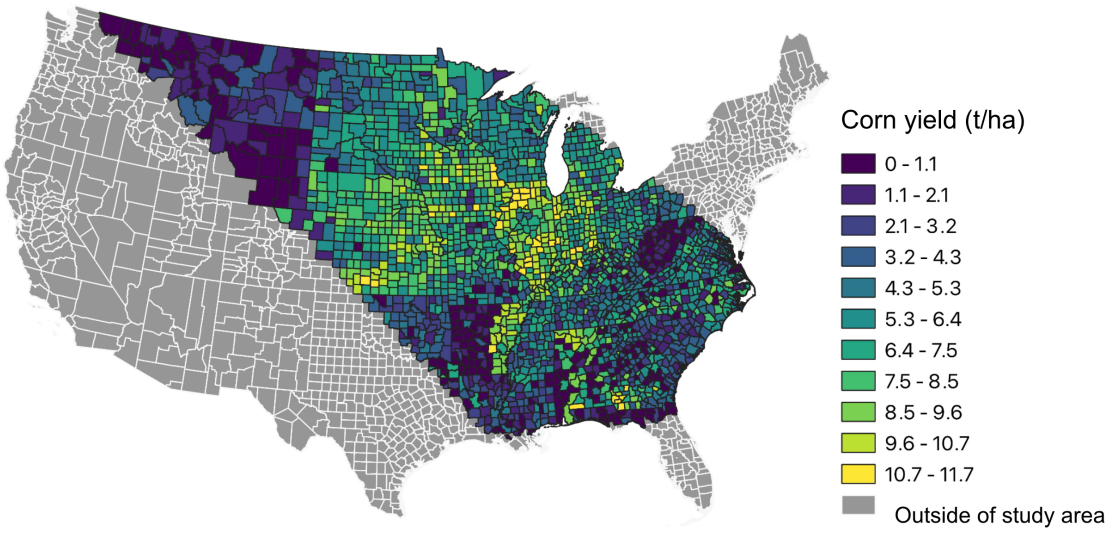

Figure 3.1: Spatial distribution of average corn yield according to the USDA NASS data for each county in t/ha from 2001-2020.

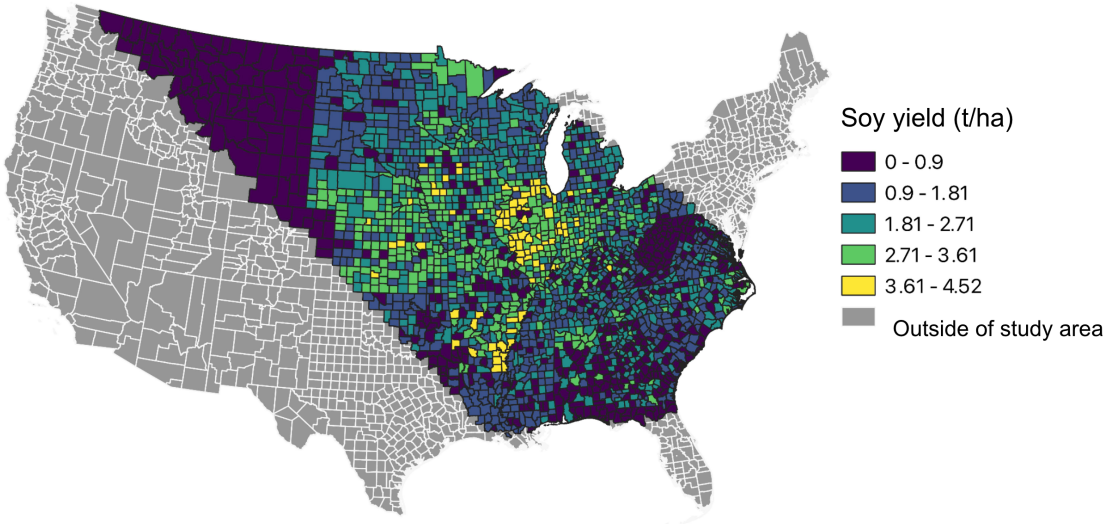

Figure 3.2: Similarly to Figure 3.1, this shows the spatial distribution of average soybean yield for each county in t/ha over the 20 year period from the NASS data.

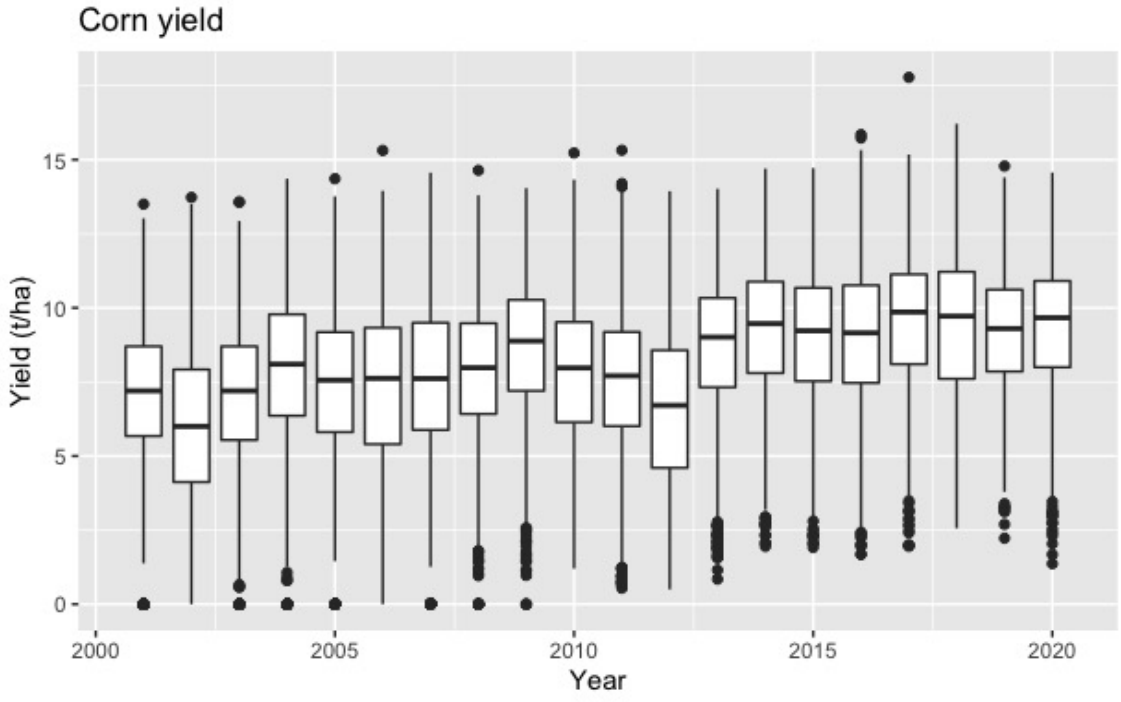

Figure 3.3: Boxplots of annual corn yields during the study period (2000 – 2020).

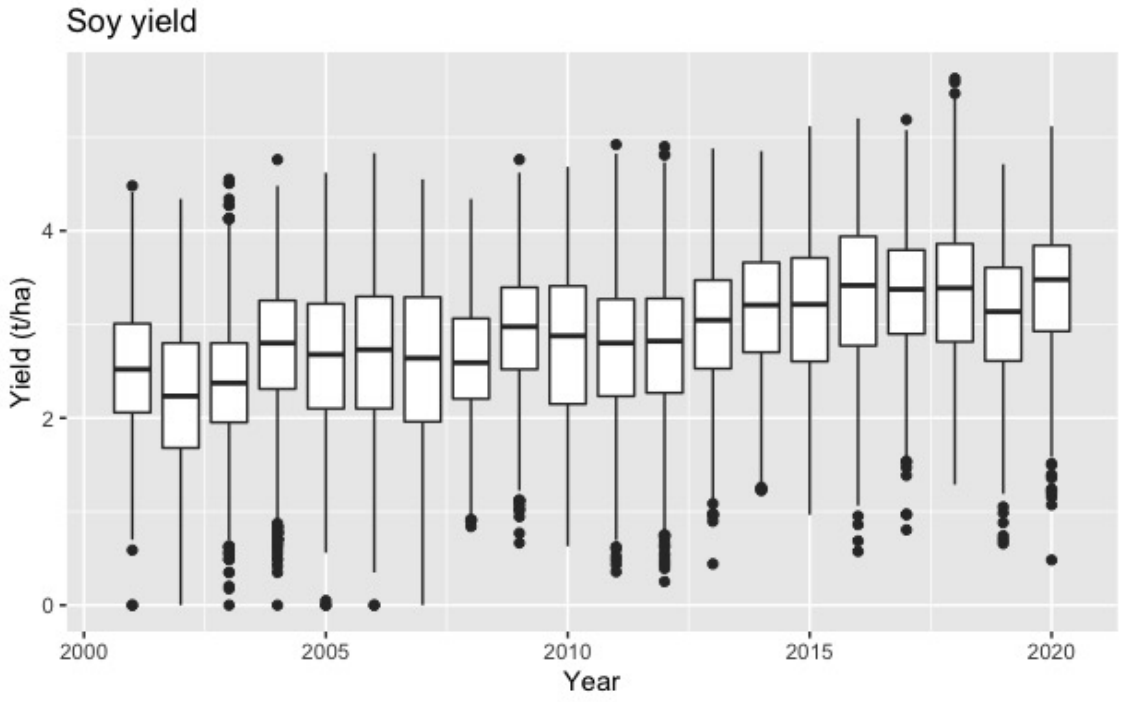

Figure 3.4: Boxplots of annual soybean yields during the study period (2000 – 2020).

3.4 Data

3.4.1 Reference data

The US Dept. of Agriculture (USDA) National Agricultural Statistics Service (NASS) provides data on annual crop production (USDA NASS, 2023). For this study I use the county-level yield data for corn and soybean in 18 states covered by MODIS tiles h10v04, h10v05, h11v04, and h11v05 (Figures 3.1 and 3.2). The Cropland Data Layer (CDL) (Han et al., 2012; Xian et al., 2009) is the USDA annual crop map. From 2001-2007, the CDL is only available for a handful of states. From 2001 - 2006 the CDL is available at 30 m spatial resolution, in 2007 it is provided at 56 m resolution, and back to 30 m in 2008. In years and counties where a crop mask is not available from the CDL, I use the crop mask developed in David Lobell's research group at Stanford University (Wang et al., 2020b). There are slight differences between these crop masks, in particular, early versions of the mask may have anomalous classifications within clearly defined fields, due to the uncertainties in crop classification, while later versions of the mask apply additional mode filtering to correct for these erroneously classified pixels. However, I am aggregating the CDL or Lobell group mask up to the size of a 500 m MODIS sinusoidal grid cell, and these small errors do not appear at that spatial scale.

3.4.2 Remotely sensed indicator data

For the SR and VI inputs I use the nadir bidirectional reflectance distribution function (BRDF) adjusted reflectance (NBAR) MODIS product, MCD43A4 (Wang et al., 2018; Schaaf et al., 2002). MCD43A4 is a combined Terra and Aqua, daily, 500 m resolution, tiled MODIS product, which gives the BRDF corrected reflectance for each of the 7 MODIS land bands (R, NIR, B, G, and 3 SWIR bands). It is important to use BRDF corrected reflectance because Terra and Aqua have different overpass times, meaning that each observation will be illuminated by the sun from different directions, which when not corrected for BRDF will result in large discrepancies between observations.

The Normalized Difference Vegetation Index (NDVI) is one of the most commonly used vegetation indices. However, when reflectance in the NIR band is very high, which is common for soybean, NDVI can saturate quickly. In order to account for the quick saturation of NDVI, instead I use the enhanced vegetation index (EVI). I calculated EVI from NBAR as follows:

$$EVI = Gf \frac{NIR - R}{NIR + C_1 R - C_2 B + L} \tag{3.1}$$

Using a gain factor, Gf, of 2.5, red and blue band aerosol resistance coefficients, C_1 and C_2, of 6 and 7.5 respectively, and a canopy background adjustment factor, L, of 1 according to Huete et al. (1994, 1997, 2002).

The MCD18A2 collection 6 (C6) product (Wang et al., 2020a) is a 3-hourly,

1 km resolution, tiled MODIS product of total PAR. The product also contains instantaneous estimates of total, direct, and diffuse PAR. The product is derived using a look-up-table approach calculating the radiative transfer inversion offline, and takes as its observational inputs, top-of-atmosphere radiance, viewing geometry, atmospheric water vapor, and aerosol optical depth. When observational data is unavailable, the MCD18A2 product takes reanalysis and climatological data as input. The direct and diffuse components are estimated by the empirical partitioning of direct and diffuse radiation due to atmospheric water vapor and aerosols. For this work, I used an internal beta version of the official product due to delays in the NASA data processing queue.

3.5 Methods

3.5.1 Data preparation

The CDL must be reprojected from Albers Conical Equal Area projection to the sinusoidal MODIS grid, then aggregated from 30 m to 1 km (actually 926.625433004 m) pixels. Using the CDL reprojected and aggregated to 1 km, all pixels for the corresponding crop type in each county were averaged to yield a single reflectance per band and PAR value for each county. Crop maps were extracted from the reprojected CDL at 30 m scale and aggregated up to the 1 km scale in the MODIS projection. Pixel purity percent was calculated according to the following:

$$P_C^r = \sum \frac{C_{30}^r}{T_{30}^r} \times 100, \tag{3.2}$$

where P_C^r is the percent of 30 m pixels of a given crop type C in an aggregated pixel of resolution r, and T_{30}^r is the total number of 30 m pixels within the aggregated pixel. For the analysis, I used only pixels with a pixel purity of 95% or higher to limit uncertainty due to mixed pixels. Because I used an internal beta version of MCD18A2(C6), I have no metadata that indicates whether the values of a given pixel are taken from observations or climatology data, so all pixels with the appropriate pixel purity are included. This should be taken into consideration as a caveat to the work presented.

3.5.2 Crop yield modeling

Physical based LUE models are based on a linear function relating GPP to the amount of PAR a leaf (or scaled up to the total canopy level) absorbs, $aPAR$, by a LUE coefficient:

$$GPP = LUE \times aPAR. \tag{3.3}$$

This relationship can also be expressed as the LUE coefficient times the total PAR available, times the fraction of PAR ($fPAR$) absorbed by the leaf or canopy:

$$GPP = LUE \times PAR \times fPAR. \tag{3.4}$$

For this work, I use VI as a proxy for $fPAR$, in order to make easier cross comparisons between my findings and other studies such as Johnson (2016) and Skakun et al. (2021). This is possible because of the correlation between $fPAR$ and

VI, and give the equation:

$$GPP = m \times PAR \times VI + b, \tag{3.5}$$

where m is the regression slope, and is b is the intercept. In this case, m should vary with crop type as it is essentially the LUE.

The basis for my crop yield model begins with Johnson (2016) and Skakun et al. (2021), where yield is correlated with the VI of a crop type or multiple SR. In order to add PAR into the model, I take two approaches. First, adding PAR as a simple variable into the multiple linear regression of VI or SR to yield, and second, I make use of light use efficiency (Eq. 3.5), where GPP is used as a proxy for yield.

For my yield models, I consider the multivariate regression linear regression the different possible combinations of PAR, SR, and VI with inputs as follows:

Input features	**Details**	**No. inputs**
PAR + EVI	accumulated daily MODIS PAR, daily EVI	2
PAR + SR	MODIS PAR and daily Red, Green, Blue, NIR BRDF corrected bands	5
PAR × EVI	accumulated daily MODIS PAR, daily EVI	1
PAR only	accumulated daily MODIS PAR	1
EVI only	daily EVI	1
SR only	daily BRDF corrected Red, Green, Blue, NIR bands	4

Table 3.1: Input features and description for each of the yield models tested.

3.5.3 Temporal analysis

First I train an annual temporal model, where for each day of the year, I calculate the regression between yield and my inputs shown in Table 3.1 and record

the R^2_{adj}. PAR is accumulated from DOY 100 in early April to DOY 225 (mid-August) for corn and DOY 255 (mid-September) for soybeans. These dates were chosen for the *a priori* knowlege of the typical peak VI for corn and soybean (around DOY 200 and 220 respectively). I use all counties as inputs for each year and crop type, which gives me 600-1000 training samples to use per year. In this way, I generate an annual temporal curve that allows me to find the day of year (DOY) most strongly correlated with yield for each year and crop type from which I then calculate and report the mean DOY of max R^2_{adj}.

To validate the models, I use the leave-one-out approach; that is I train the model on 19 of the 20 years of available data, withholding a single year for validation and iterate through all 20 years. This way I can see both the inter-annual variability and the overall average results of the models.

3.6 Results

3.6.1 Corn

Prior to evaluating the model results, I check my *a priori* assumptions by comparing the temporal correlation coefficient between county yield and daily EVI, in accordance with Johnson (2016), and find good agreement with the literature. The temporal correlation between EVI and county yield for corn is shown in Figure 3.5 and agrees with the findings of Johnson (2016), and gives confidence for the *a priori* assumptions previously described.

The best performing models for corn yield were the PAR + VI and PAR + SR

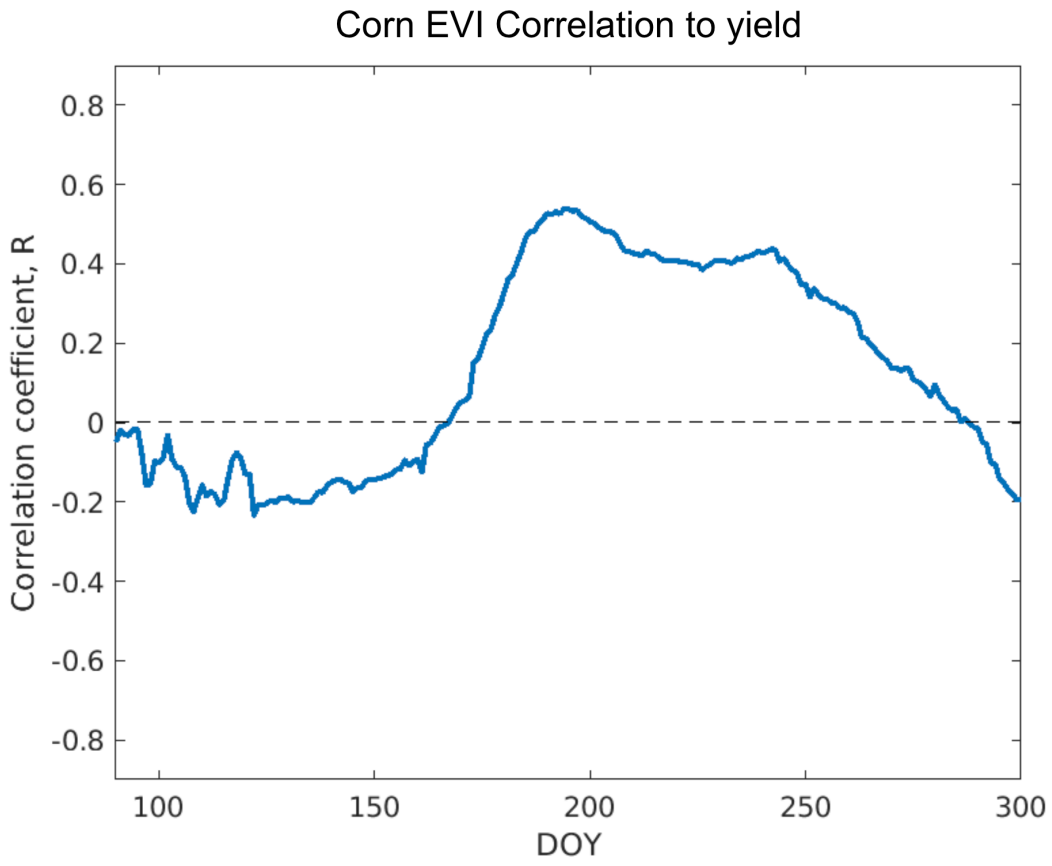

Figure 3.5: Correlation coefficient between EVI and county crop yield for all counties in the study area.

models, with an R^2_{adj} of 0.76 and 0.81 respectively shown in Figure 3.6 and Table 3.2. All models show the highest R^2_{adj} at around DOY 190 in early July. The VI and SR only models reach a maximum R^2_{adj} of 0.65 and 0.67 respectively, which is aligned with the findings of Johnson (2016), and the best in field results from Skakun et al. (2021). The PAR only and PAR×VI model reach a maximum R^2_{adj} of 0.37 and 0.41 respectively, indicating that PAR alone is not a sufficient indicator of yield.

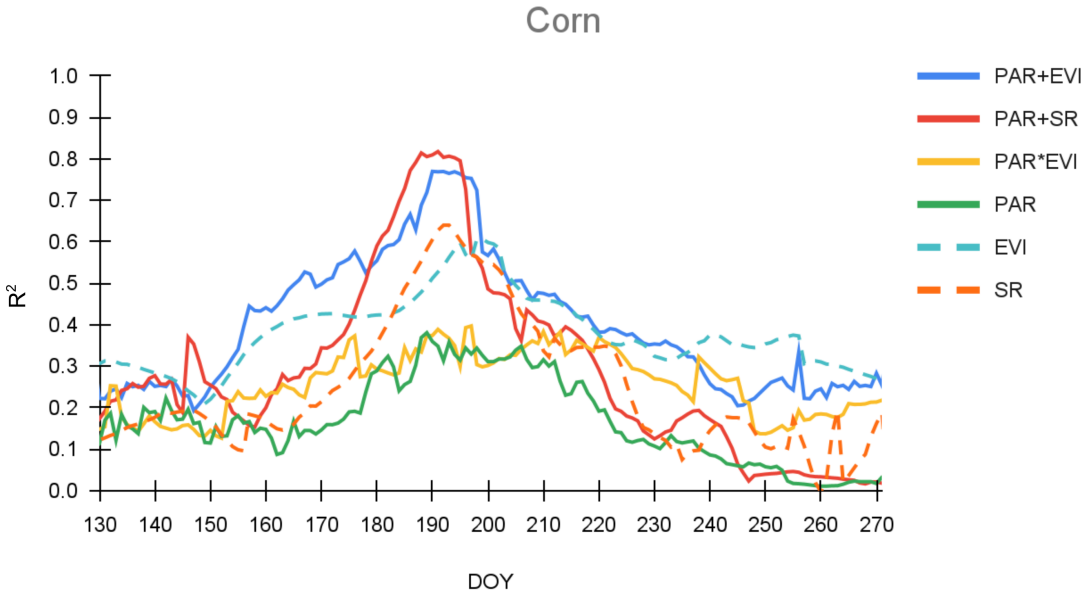

Figure 3.6: Averaged temporal R^2_{adj} results of the six different models.

For the best performing model, PAR + SR, I show the scatter plot of estimated yield compared to observed yield in t/ha in Figure 3.7. The model underestimates yield and has a root mean square error (RMSE) of 6%. The majority of the observations are around 10-11 t/ha, while the model estimates yield at 8-9 t/ha. I also show the trend in R^2 over time in Figure 3.8. There is a slight decreasing trend in R^2 over time that coincides with the increasing trends in crop yield (Figure 1.3.

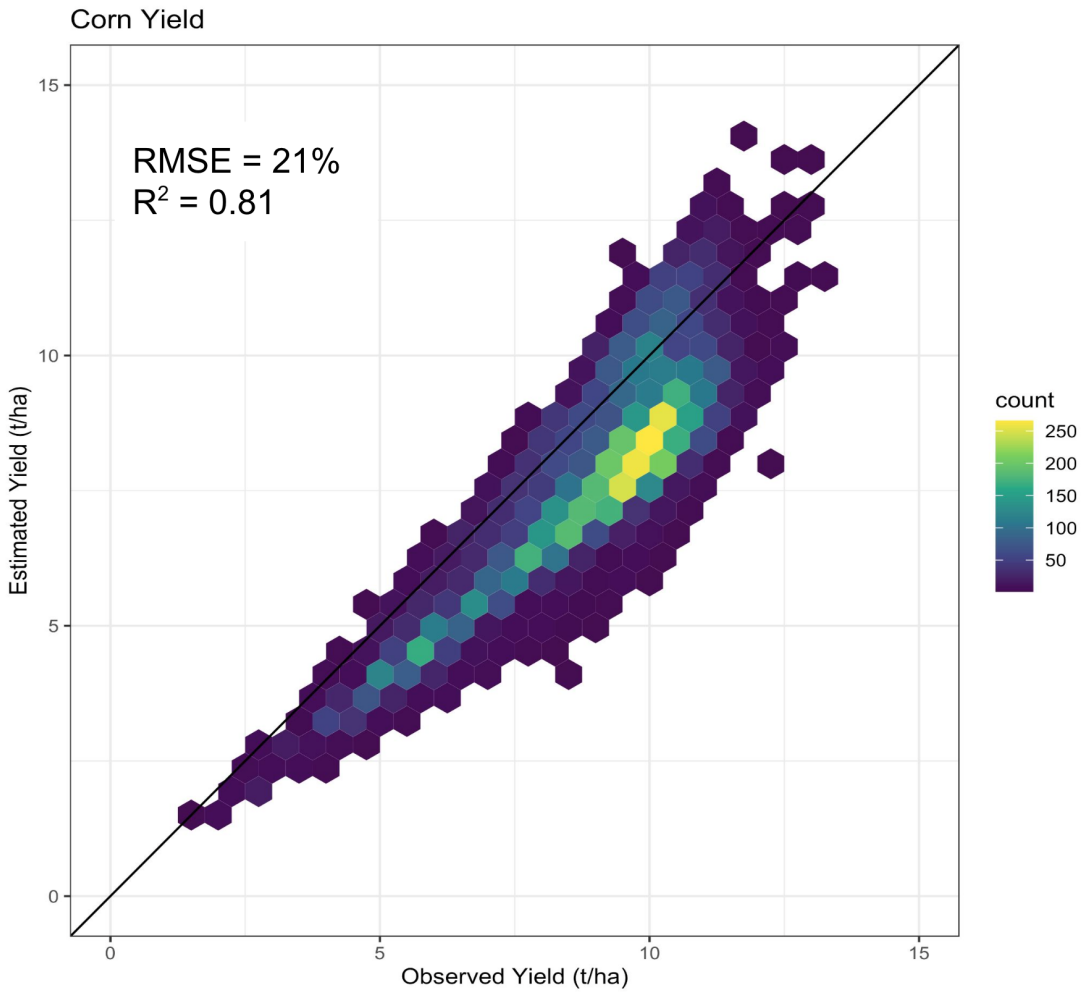

Figure 3.7: Modeled vs observed county yields for the PAR + SR model for corn.

Year	PAR + EVI	PAR + SR	PAR × EVI	PAR only	EVI only	SR only
2001	0.80	0.87	0.46	0.32	0.75	0.85
2002	0.73	0.75	0.41	0.31	0.64	0.72
2003	0.85	0.89	0.45	0.45	0.73	0.73
2004	0.84	0.88	0.41	0.40	0.71	0.73
2005	0.83	0.89	0.47	0.38	0.74	0.75
2006	0.79	0.87	0.34	0.41	0.64	0.66
2007	0.83	0.87	0.47	0.37	0.72	0.73
2008	0.77	0.82	0.35	0.39	0.63	0.66
2009	0.68	0.76	0.31	0.34	0.56	0.64
2010	0.79	0.85	0.52	0.36	0.69	0.62
2011	0.73	0.76	0.43	0.34	0.61	0.60
2012	0.79	0.80	0.40	0.40	0.59	0.65
2013	0.84	0.87	0.44	0.41	0.73	0.64
2014	0.61	0.65	0.31	0.30	0.52	0.63
2015	0.70	0.77	0.41	0.32	0.63	0.63
2016	0.72	0.77	0.47	0.31	0.56	0.66
2017	0.75	0.82	0.34	0.45	0.69	0.64
2018	0.73	0.76	0.47	0.40	0.61	0.62
2019	0.74	0.75	0.35	0.38	0.59	0.60
2020	0.68	0.76	0.31	0.41	0.73	0.65
Avg	**0.76**	**0.81**	**0.41**	**0.37**	**0.65**	**0.67**
Std	**0.06**	**0.06**	**0.06**	**0.05**	**0.07**	**0.06**

Table 3.2: Validation R^2_{adj} of corn models for each year, by input features, with average and standard deviation reported in bold.

Field studies, such as Skakun et al. (2021) show a similar result, that is when yields are highest, R^2 tend to be lower, this is due to the limitation of spectral remote sensing to detect all influences on crop yield.

Year	PAR + EVI	PAR + SR	PAR × EVI	PAR only	EVI only	SR only
2001	189	189	196	191	195	194
2002	176	174	183	184	178	174
2003	195	192	203	194	200	195
2004	193	190	206	196	198	192
2005	178	174	186	174	183	175
2006	183	177	195	178	197	177
2007	173	172	171	170	178	172
2008	194	190	211	197	202	206
2009	190	192	205	197	197	192
2010	202	195	201	209	203	203
2011	191	191	206	209	186	187
2012	179	175	192	185	193	181
2013	178	174	186	182	185	175
2014	183	185	183	188	182	184
2015	170	187	185	184	185	173
2016	190	205	195	190	194	195
2017	197	193	187	197	197	195
2018	201	199	205	195	205	205
2019	205	200	199	202	190	206
2020	199	203	203	203	205	205
Avg	**188**	**188**	**195**	**191**	**193**	**189**

Table 3.3: DOY of maximum corn R^2_{adj} for each year as well as the 20 year average.

3.6.2 Soybeans

Again, prior to evaluating model performance, I check my *a priori* assumptions for soybean by comparing the temporal correlation coefficient between EVI and yield for soybeans in my study area and find agreement with Johnson (2016), shown in Figure 3.9.

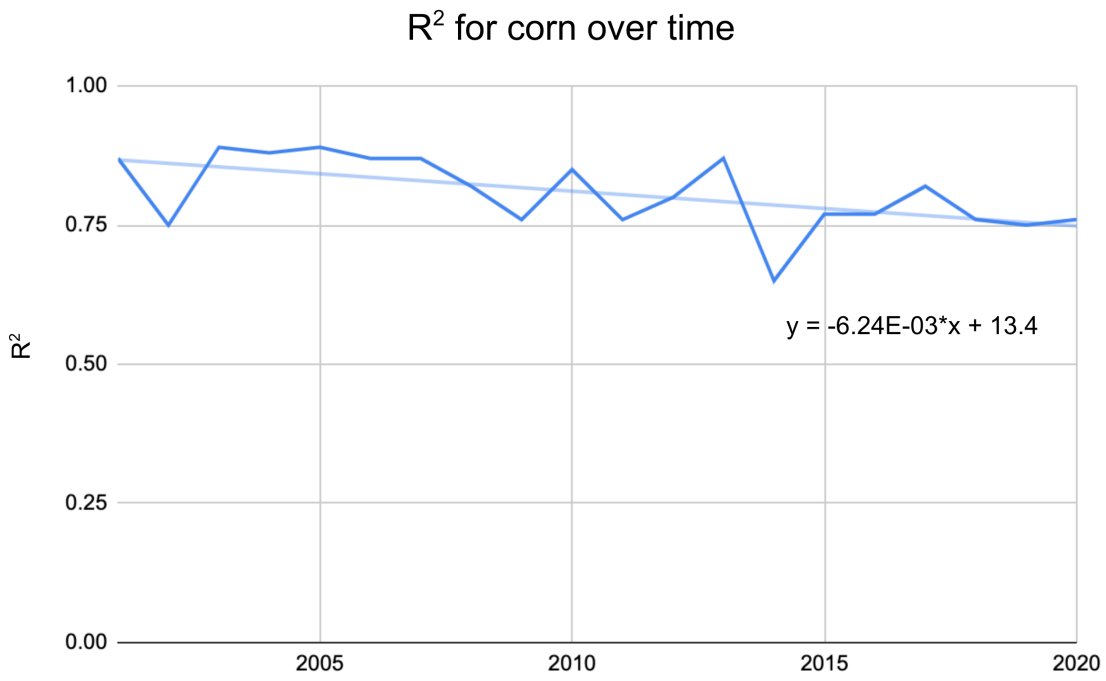

Figure 3.8: Trend in R^2 for corn over time. As yield increases, R^2 decreases slightly and this coincides with increased trends in crop yield.

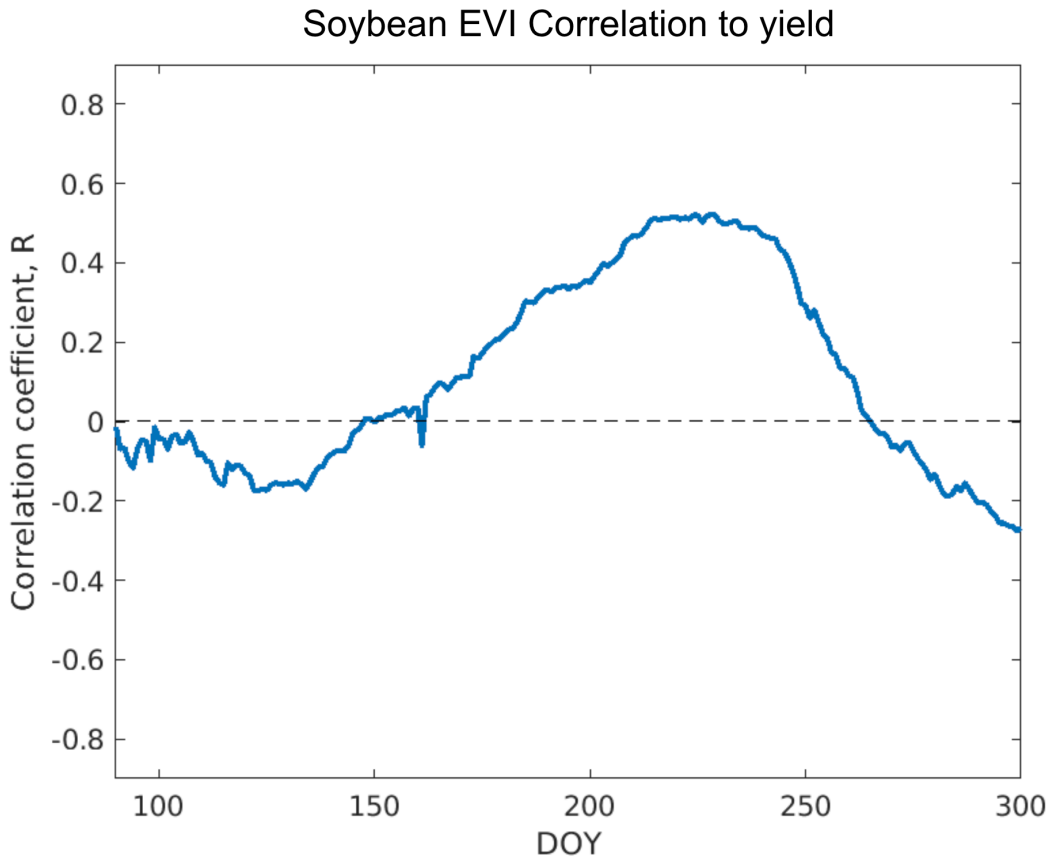

Figure 3.9: Temporal correlation coefficient between EVI and soybean yield for all counties in the study area.

The model results for soybean are shown in Figure 3.10 and Tables 3.4 and 3.5. The best performing models were PAR + VI and PAR + SR, with an R^2_{adj} of 0.78 and 0.80 respectively. The VI and SR only models give a maximum R^2_{adj} of 0.59 and 0.57 respectively, again in line with Johnson (2016).

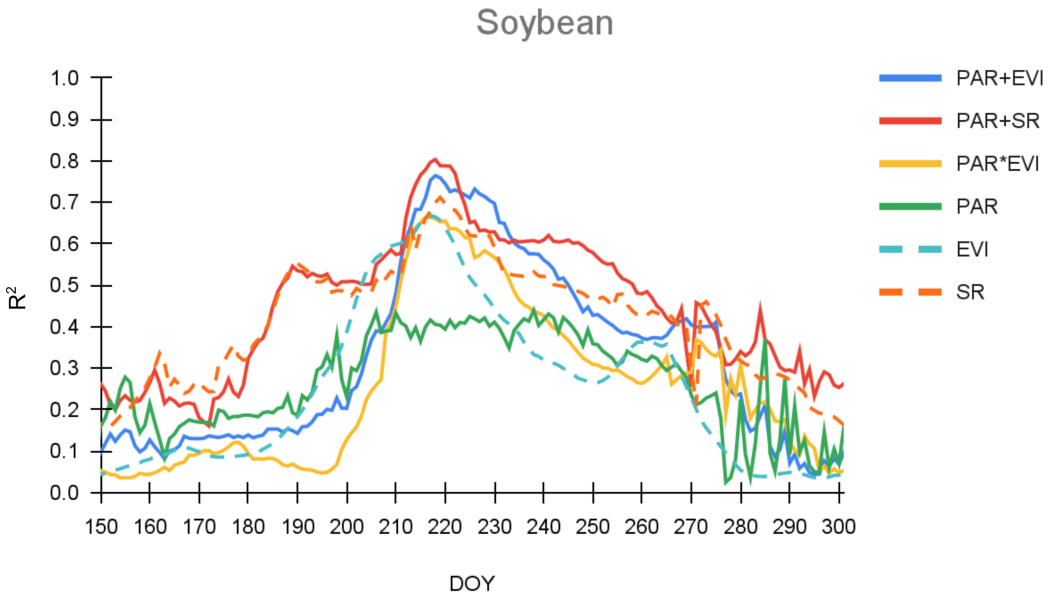

Figure 3.10: Averaged temporal validation results for the six soy models.

Figure 3.11 shows the scatter plot of estimated yield to observed yield for the best performing model, PAR + SR. Again, the model underestimates yield and similarly has an RMSE of 9%, however, for soybean the underestimation of yield is much smaller than for corn. The majority of the observations and modeled soybean yields are between 2 and 3 t/ha. Again, I show the trend in R^2 over time in Figure 3.12, and again I see a slight decrease in R^2 that coincides with the increases in soybean yield seen in Figure 1.3.

Year	PAR + EVI	PAR + SR	PAR × EVI	PAR only	EVI only	SR only
2001	0.68	0.81	0.80	0.44	0.67	0.67
2002	0.94	0.96	0.77	0.40	0.66	0.70
2003	0.85	0.85	0.55	0.44	0.52	0.50
2004	0.96	0.97	0.69	0.47	0.59	0.57
2005	0.92	0.92	0.89	0.40	0.76	0.70
2006	0.81	0.84	0.61	0.38	0.53	0.52
2007	0.77	0.78	0.65	0.42	0.66	0.65
2008	0.77	0.85	0.69	0.46	0.56	0.55
2009	0.67	0.68	0.50	0.46	0.50	0.48
2010	0.65	0.79	0.56	0.45	0.55	0.55
2011	0.74	0.75	0.61	0.42	0.62	0.59
2012	0.88	0.90	0.70	0.41	0.70	0.63
2013	0.77	0.77	0.60	0.40	0.62	0.60
2014	0.64	0.65	0.51	0.42	0.52	0.50
2015	0.69	0.70	0.56	0.44	0.54	0.55
2016	0.72	0.73	0.57	0.46	0.57	0.53
2017	0.69	0.70	0.50	0.44	0.50	0.48
2018	0.75	0.77	0.54	0.43	0.55	0.51
2019	0.86	0.88	0.61	0.39	0.57	0.60
2020	0.75	0.76	0.55	0.42	0.54	0.52
Avg	**0.78**	**0.80**	**0.62**	**0.43**	**0.59**	**0.57**
Std	**0.10**	**0.09**	**0.11**	**0.03**	**0.07**	**0.07**

Table 3.4: Validation results for the soy models for each year and the 20-year average and standard deviation, by input features.

Year	PAR + EVI	PAR + SR	PAR × EVI	PAR only	EVI only	SR only
2001	221	217	221	210	222	218
2002	224	222	226	219	217	222
2003	200	225	195	200	196	199
2004	219	220	204	226	205	220
2005	211	217	211	219	215	218
2006	225	227	211	217	217	218
2007	219	219	209	217	221	219
2008	215	211	207	209	213	212
2009	225	227	192	204	215	225
2010	214	195	214	194	217	218
2011	210	200	204	200	209	202
2012	200	215	215	200	216	219
2013	214	212	216	219	217	212
2014	220	223	217	220	220	223
2015	211	217	211	219	215	218
2016	216	223	210	203	213	223
2017	215	215	217	207	215	212
2018	230	232	219	220	229	233
2019	215	211	223	219	225	218
2020	215	226	217	213	210	215
Avg	**216**	**218**	**212**	**212**	**215**	**217**

Table 3.5: DOY of maximum soy validation R^2_{adj} for each year and the 20-year average.

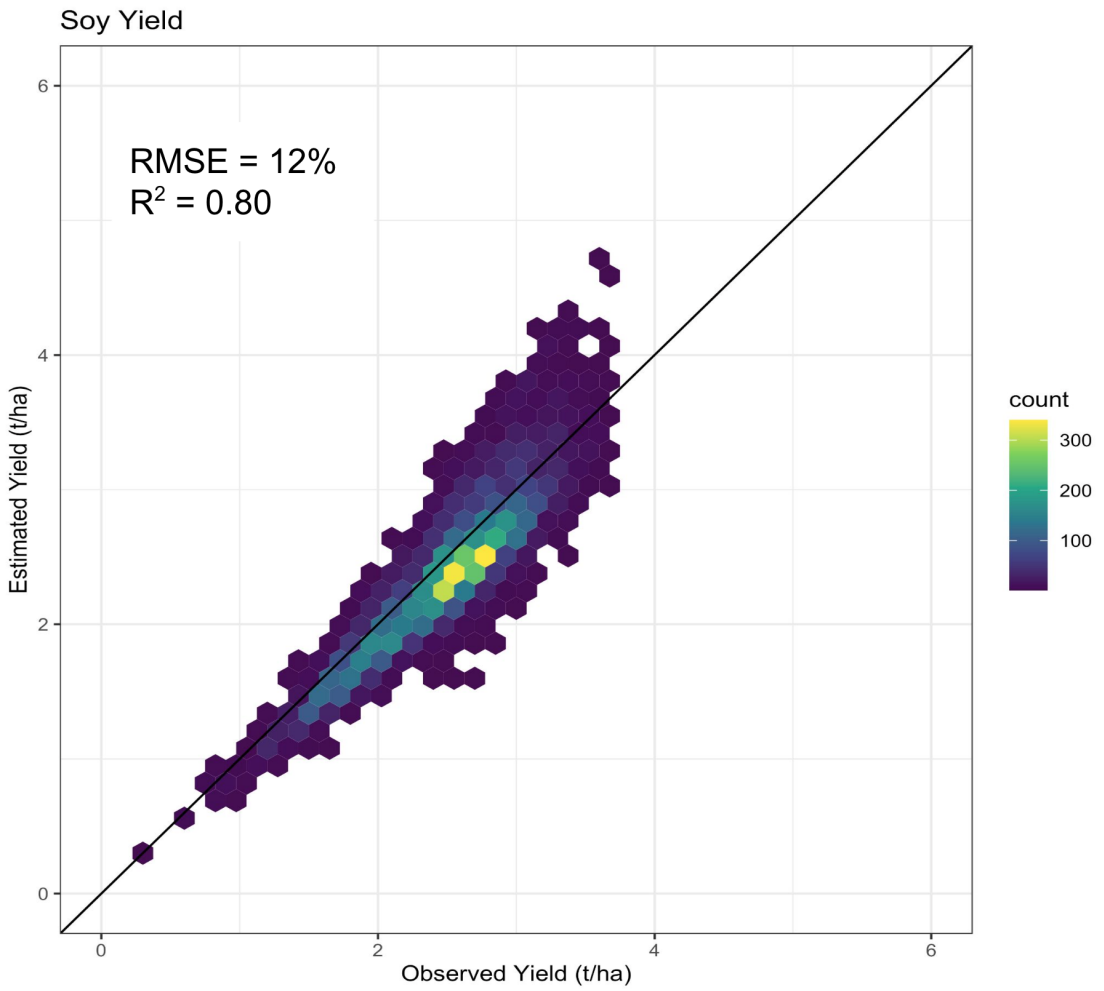

Figure 3.11: Modeled vs observed county yields for the PAR + SR model for soybean.

3.7 Discussion and Conclusions

In this chapter, I incorporated PAR as an explicit indicator into empirical crop yield models of corn and soy at the county level in the United States from 2001-2020. My empirical models are based on those described in Johnson (2016) and Skakun et al. (2021), which make use of vegetation indices and surface reflectance as indicators. While the spectral response of vegetation implicitly includes the information about PAR that the canopy is receiving, explicitly adding it to the

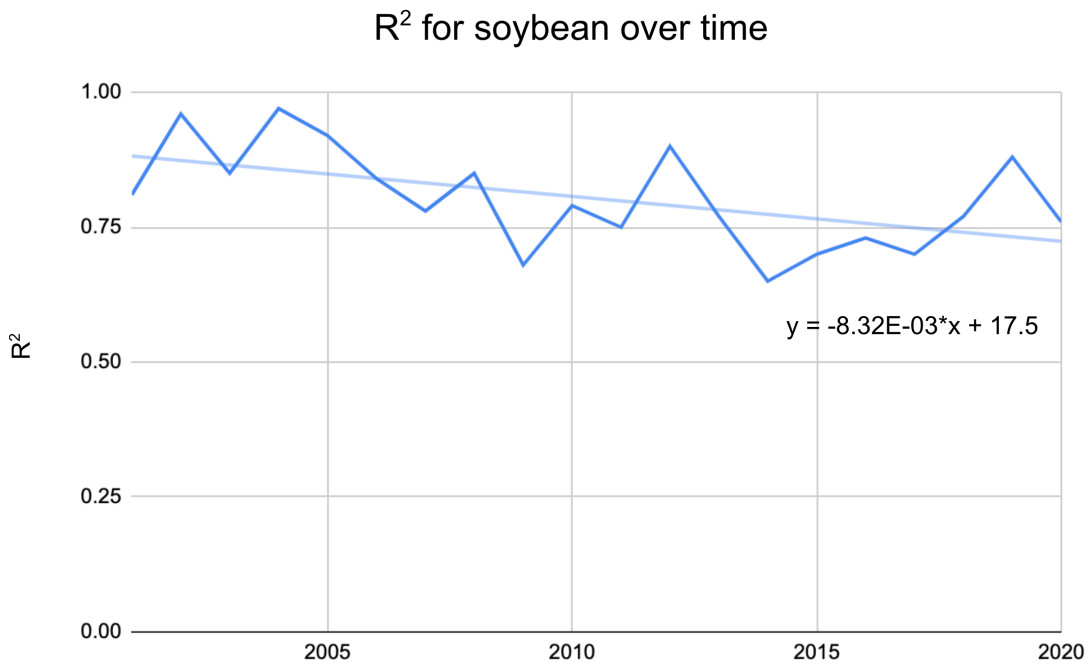

Figure 3.12: Trend in R^2 for soybean over time. As yield increases, R^2 decreases slightly, coinciding with increases in soybean yield.

model increases the yield variance explained by the model by up to 20%, and shows earlier indication of crop yield compared to single indicator models.

One limitation to this study is the strict dates between which PAR is accumulating. Future efforts should be more sensitive to the real-time planting and emergence of crops, however for this study such detailed data for all pixels in the study area was not available. As yield for both corn and soy increase, the variability that is explained in these models decreases slightly, suggesting that the increases in yield are not to do with changes in PAR, which agrees with studies such as Wild et al. (2013) that show how solar radiation over the US has not been trending since the 1970s and 1980s due to the Clean Air Act.

Prior to the MODIS PAR product, MCD18A2 Wang et al. (2020a), global, tiled

or gridded, PAR retreivals were not as widely accessible for addition into empirical crop yield models, like those I presented here. Other existing products, such as CERES or GLASS, were available at coarser spatial resolutions that would not have been suitable for yield modeling, even when aggregated to the county scale. The results of my county-level yield modeling align with field-level studies such as (Johnson, 2016) and (Franch et al., 2019). Specifically, for both corn and soybean, the VI-only model shows a maximum R^2_{adj} of around 0.60, which corresponds very well to (Johnson, 2016)'s findings, and my results explicitly including PAR show maximum R^2_{adj} of 0.81 and 0.80 for corn and soybean respectively. These results complement the seminal research on VI-modeling of vegetation.

I have shown that the addition of PAR to the model improves on a SR or VI based models, even at MODIS scale and in the United States where radiation is not the limiting factor for vegetation growth (Milesi et al., 2005). Given this modest improvement, in the United States, where agricultural systems are highly managed and radiation is not the limiting factor, I assert that my findings indicate that the addition of PAR to empirical yield models in regions in the world where vegetation is radiation limited may add significant information. However, a limiting factor to that work is the spatial scale of PAR data. I suggest that the radiation community may be able to add significant value to the crop yield modeling community by calculating or super resolving surface radiation to finer spatial resolution satellite data such as from Harmonized Landsat-Sentinel, Planet, RapidEye, or WorldView. Additionally, I consider this work to be an indirect assessment of the MODIS PAR (MCD18A2) product, since it added value to these crop yield models. I find that the MCD18A2

product is very useful for yield modeling studies, however, it should be noted that without an indication of which pixels have climatological data as input, there is a level of uncertainty which cannot be accounted for.

Future work should be directed at the field scale, and eventually the sub-field scale as that level is often of greater interest to farmers and government entities responsible for overseeing and managing agricultural polices and practices. This will require either super resolving the MODIS product (MCD18A2) down to a finer spatial resolution, or generating satellite derived estimates of PAR from higher resolution satellites. A variety of different methods are available for both options, however each will require additional validation before they can be incorporated into empirical crop yield models.

Chapter 4: Using the absorption coefficient of PAR to capture crop type and irrigation method for large fields

4.1 Overview

Due to the differences in biochemistry, cell structure, and photosynthetic pathways, different plant species absorb photosythetically active radiation (PAR) with varying efficiency and have evolved to thrive in different conditions, such as direct, intense sunlight or indirect, diffuse light conditions. In-field measured yield, canopy chlorophyll content (CCC), and the coefficient of absorption of PAR, (α_{PAR}), show strong, species specific relationships. In this work, I determine to what extent MODIS-derived (α_{PAR}) is suitable for capturing the same relationships. I found that MODIS-derived α_{PAR} corresponds to the plant CCC in the same manner as ground-based α_{PAR} measurements. Specifically, I found that for three experimental fields of corn and soybean fields in Eastern Nebraska R^2 was 0.97 and RMSE was 1.34 (11%) when comparing MODIS-derived α_{PAR} with the in situ measurements. I also found that the relationships between MODIS-based α_{PAR} and CCC for corn and soybean corresponded to the ones obtained from in situ data. The relationships between α_{PAR} and CCC for corn and soybean are distinct due to the different photo-

synthetic pathways of corn and soybean, differences in cell structure, and chloroplast distribution between the two crops. Crop yield and productivity are also related to CCC, meaning MODIS α_{PAR} can be used as a crop specific indicator of yield at large scale.

4.2 Introduction

Vegetation activity (photosynthesis) requires sunlight, precipitation, and favorable temperatures (Nemani et al., 2003; Running et al., 2004; Milesi et al., 2005). The more efficiently sunlight is absorbed by cropped vegetation, the higher yields can be (Gitelson et al., 2015; Yuan et al., 2016). Canopy chlorophyll content (CCC) accounts for 90% of the variation in crop yield (Gitelson et al., 2014; Peng et al., 2017). In general, crops with higher chlorophyll content tend to have higher yields, as they are able to photosynthesize more efficiently to produce more biomass.

A plant's structural, chemical, and biophysical characteristics will impact how efficiently it can convert carbon from the atmosphere to energy for the plant during photosynthesis. Corn (maize) uses the C4 photosynthetic pathway, while soybean uses the C3 pathway (Boyer, 1970). In the C4 pathway, carbon dioxide (CO_2) is transformed into a four-carbon compound (vs the three-carbon compound used by the C3 pathway), which is then transported to bundle sheath cells (which C3 plants do not have) where it releases the CO_2 used to produce the sugars the plant needs. The C4 pathway is an evolution of the C3 pathway that reduces photorespiration and enhances the efficiency of photosynthesis, allowing C4 plants to survive under

hotter and drier conditions compared to C3 plants (Ehleringer and Cerling, 2002).

The coefficient of the absorption of PAR, α_{PAR}, a measure of how efficiently and effectively a plant can absorb photosynthetically active radiation for photosynthesis, is a semi-analytically defined unitless quantity that is related to biophysical quantities such as the fraction of absorbed PAR, $FAPAR$ and canopy chlorophyll content. α_{PAR} is sensitive to plant biochemistry, structural properties, and photosynthetic pathway (Gitelson et al., 2019, 2021), and shows a stronger linear correlation to crop yield than vegetation indices (VIs), such as the normalized difference vegetation index (NDVI), or those that can better account for saturation due to dense, healthy vegetation, e.g., the enhanced vegetation index (EVI), or the wide dynamic range vegetation index (WDRVI) (Gitelson, 2004). NDVI, EVI, and WDRVI are defined as follows:

$$NDVI = \frac{NIR - Red}{NIR + Red}, \tag{4.1}$$

where NIR is the reflectance in the near infrared (841–876 nm) and Red is the reflectance in the red region (620–670 nm).

$$EVI = Gf \frac{NIR - Red}{NIR + C_1 R - C_2 B + L}, \tag{4.2}$$

where B is the surface reflectance in the blue region (∼400 - 485 nm), L is an adjustment factor for the canopy background, C_1 and C_2 are aerosol resistance coefficients, and Gf is a gain factor specific to the sensor.

$$WDRVI = \frac{a\ NIR - Red}{a\ NIR + Red}, \tag{4.3}$$

where the weight, a, can vary from 0.1 - 0.2 to account for high values of NIR reflectance in dense healthy vegetation, such as cultivated crops.

Surface radiation trends are associated with trends in both precipitation and near-surface air temperature (Wild, 2012), which impacts plant growth and crop yields. Studies have shown how human and natural activity have affected light conditions, particularly with respect to atmospheric aerosols (Roderick et al., 2001; Gu et al., 2003; Rap et al., 2015, 2018), and that with those changing light conditions the amount of carbon removed from the atmosphere during photosynthesis increased (Alton et al., 2007; Mercado et al., 2009; Kanniah et al., 2012; Cheng et al., 2015). As the planet changes due to global warming, and the various geoengineering strategies (Irvine et al., 2016; Lockley et al., 2020; Liu et al., 2021) designed to avert some of the adverse effects of climate change, it is increasingly important to be able to monitor and study our cropped vegetation explicitly including radiation as a forcing or indicator.

In this chapter, I assess the correspondence between MODIS-derived α_{PAR} and field measured α_{PAR} at three experimental sites in Eastern Nebraska. I also use the phenology of MODIS-derived α_{PAR} to map corn and soybean at two neighboring, non-irrigated cropped sites and compare my classifications with the Cropland Data Layer (CDL) (Han et al., 2012; Xian et al., 2009). Finally, I compare the magnitudes of α_{PAR} to average precipitation in the region and find correspondence between

above average rainfall and large values of α_{PAR}.

4.3 Study Area

For this study, I have selected three experimental research sites belonging to the University of Nebraska Agricultural Research and Development Center, which have been part of ongoing studies on crop and management studies since 2001 (Suyker, 2022). The sites are located in Saunders County, Nebraska, near the city of Mead, NE. Additionally, I have selected two neighboring cropped 500 m MODIS sinusoidal grid cells, also in Saunders County to examine the 20 year time series of α_{PAR}. A map of the sites from the University of Nebraska Agricultural Research and Development Center is shown in Figure 4.1.

Site 1 is irrigated with a center pivot system and is always planted with maize (corn). Site 2 is also on a center-pivot irrigation system, but it is planted on a maize-soybean rotation. Site 3 is a rainfed site, and it follows the same maize-soybean rotation as Site 2. Figure 4.2 shows the 2008 Cropland Data Layer (CDL) crop mask reprojected and aggregated to the 500 m MODIS grid (following the methods described in Chapter 3 Section 3.5.1) with the grid cells representing the experimental sites marked with red dots, and the cropped grid cells for the time series analysis marked with black dots.

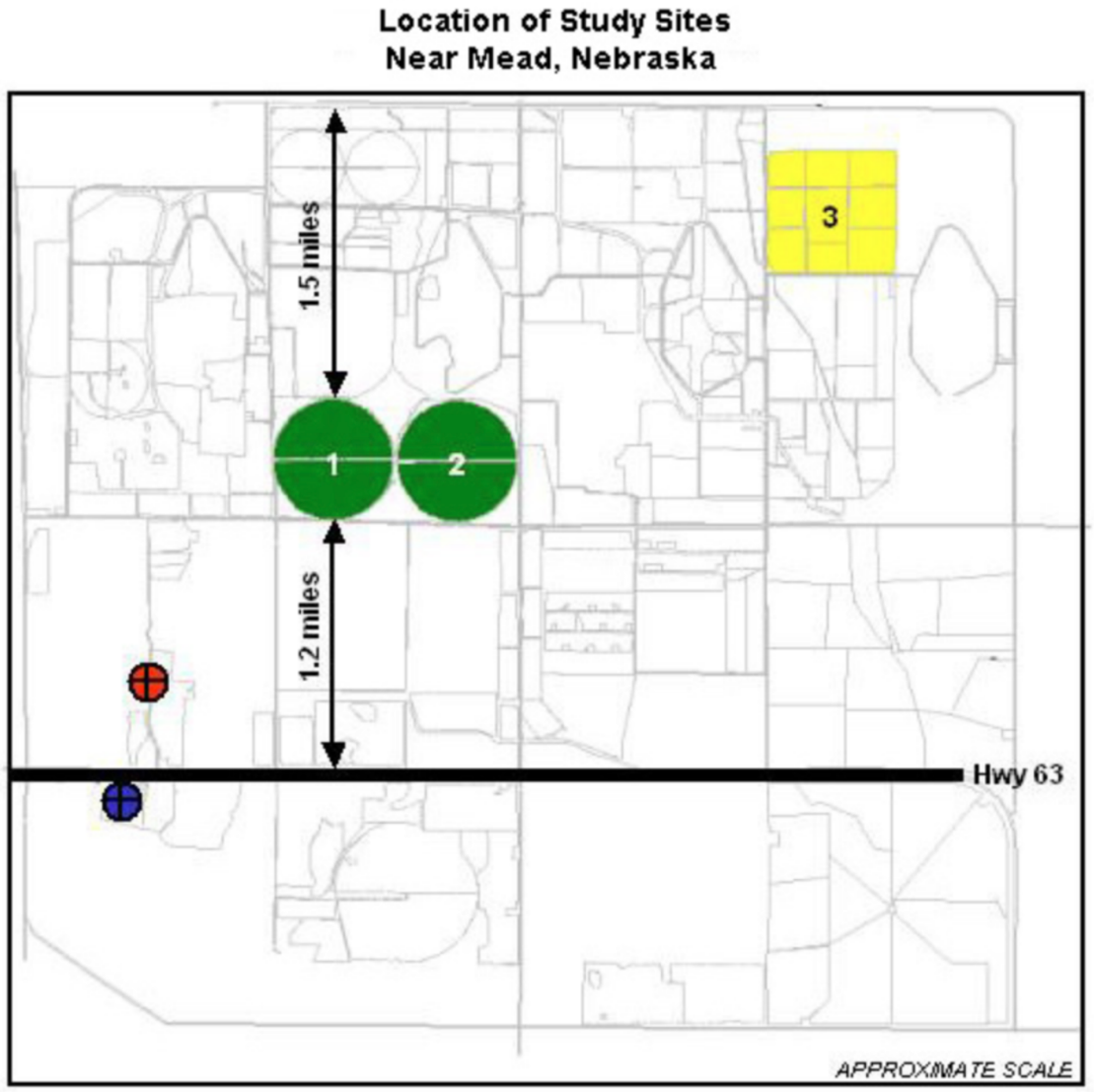

Figure 4.1: Location of the three field sites near Mead, Nebraska. Figure is from the University of Nebraska Carbon Sequestration Program (`http://csp.unl.edu/public/sites.htm`)

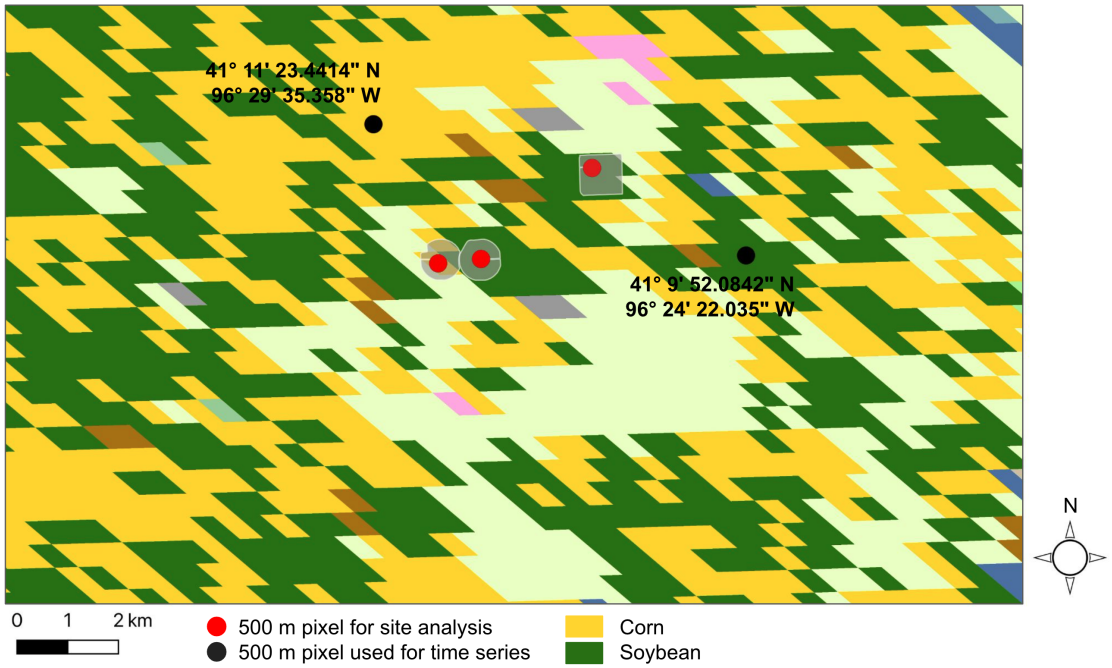

Figure 4.2: Location of the three experimental sites with the grid cell designated for analysis in red. Grid cells designated with a black marker are used for further temporal analysis of α_{PAR}.

4.4 Data

4.4.1 Ground measurements

The experimental research sites contain eddy covariance flux towers which measure CO_2, water, and energy fluxes from the canopy, as well as two hyperspectral radiometers mounted above the canopy and a set of upwelling and downwelling broadband sensors are mounted centrally on the same sensor platform. These instruments are used along with destructive sampling techniques to obtain data on biomass, canopy chlorophyll measurements, spectral measures of the canopy, including leaf area index (LAI), enhanced vegetation index (EVI), and the absorption coefficient of PAR (α_{PAR}). Additionally, each field has a record of crop type, irriga-

Field	Crop	Irrigation	Years
1	Corn (Maize)	Irrigated	2005
2	Corn (Maize)	Irrigated	2003, 2005
2	Soybean	Irrigated	2002, 2004
3	Corn (Maize)	Rainfed	2003, 2005
3	Soybean	Rainfed	2002, 2004

Table 4.1: Crop rotation for the three sites in Nebraska from 2002-2005. Field locations are shown in Fig. 4.1

tion, and other management information. These data are available from 2002-2005, and the crop rotation of the three sites is shown in Table 4.1. Additionally, yield data for these three fields is available for the years 2002-2003.

4.4.2 Remote sensing data

The Moderate Resolution Imaging Spectroradiometer (MODIS) gridded surface reflectance product (MOD09A1 C6 from Terra observations), is available at 500 m in the Red (620–670 nm), NIR (841–876 nm), Blue (459–479 nm), and Green (545–565 nm) spectral bands. Vermote et al. (2009); Bréon and Vermote (2012) developed an alternative bi-directional reflectance distribution (BRDF) correction method to the MCD43 product (Schaaf et al., 2002). Each of these sites (shown in Fig. 4.1) contains the majority of at least one 500 m MODIS grid cell, therefore I selected the Vermote et al. (2009) method BRDF corrected surface reflectance from 2002 to 2005 to calculate the coefficient of the absorption of PAR, α_{PAR}, as defined by Gitelson et al. (2021) in Eq. 4.4 to compare to the ground measurements. The cropped grid cells indicated by black dots in Figure 4.2 were selected using the CDL (Han et al., 2012; Xian et al., 2009) reprojected to the MODIS grid and aggregated to 500 m (following the method described in Chapter 3 Section 3.5.1).

4.5 Methods

The coefficient of absorption of PAR, α_{PAR}, is derived from plant reflectance spectra and is defined according to Gitelson et al. (2019) as follows:

$$\alpha_{PAR} = \frac{\rho_{NIR}}{\rho_{VIS}} - 1 \tag{4.4}$$

where reflectance in the visible spectrum, ρ_{VIS}, is equal to the mean of the reflectance in the red, green, and blue bands:

$$\rho_{VIS} = \frac{\rho_{Red} + \rho_{Green} + \rho_{Blue}}{3} \tag{4.5}$$

Using BRDF corrected MODIS surface reflectance in the visible and NIR bands at the 500 m MODIS sinusoidal grid scale (actually 463.312716502 m), I calculated α_{PAR} for the MODIS observation containing the majority of each experimental field site. Then I calculated the linear regression between field-measured and MODIS-derived α_{PAR}, recording the R^2 and RMSE for each of the three experimental sites.

$$R^2 = \frac{\sum \left(\alpha_{PAR}^{MOD} - \overline{\alpha_{PAR}^{fld}}\right)^2}{\sum \left(\alpha_{PAR}^{fld} - \overline{\alpha_{PAR}^{fld}}\right)^2} \tag{4.6}$$

$$RMSE = \sqrt{\frac{\sum \left(\alpha_{PAR}^{MOD} - \alpha_{PAR}^{fld}\right)^2}{n}} \tag{4.7}$$

where α_{PAR}^{MOD} is the MODIS-derived α_{PAR}, α_{PAR}^{fld} is the field-measured α_{PAR}, and n is the number of observations.

I also calculate the linear regression between field measured α_{PAR} and plant chlorophyll content, recording the regression slope for each crop type, and compared it to the linear regression between MODIS-derived α_{PAR} and plant chlorophyll content for the different crops in each field. In order to verify that α_{PAR} shows a distinct phenology for each crop type, I plotted α_{PAR} for each DOY for each field over the 3 years of available field data.

Finally, I calculated α_{PAR} for the entire MODIS tile (h10v04) containing the study area, selected two cropped grid cells from Saunders County, Nebraska and plotted the time series of α_{PAR} from 2000-2022 for both grid cells in order to identify the crop type and irrigation or other management information. I selected the two grid cells based on their grid cell purity, calculated using the reprojected CDL at 30 m scale aggregated up to 500 m in the MODIS projection. grid cell purity percent was calculated as follows:

$$P_C^{500} = \sum \frac{C_{30}^{500}}{T_{30}^{500}} \times 100 \tag{4.8}$$

where P_C^{500} is the percent of 30 m pixels of a given crop type C in an aggregated grid cell of resolution 500 m and T_{30}^{500} is the total number of 30 m pixels within the aggregated grid cell. For this analysis, as before, I used only grid cells with a pixel purity of 95% or higher due to the uncertainty related to mixed pixels.

4.6 Results

4.6.1 Site Analysis

MODIS-derived α_{PAR} closely matches the field measured α_{PAR}, with an R^2 of 0.97 and an RMSE of 1.34 (11%), shown in Figure 4.3, indicating the suitability of using MODIS-derived α_{PAR} for crop yield studies of large fields or large aggregated cropped areas. The relationship between field measured α_{PAR} and plant chlorophyll content (Figure 4.4) shows two distinct relationships, one for each crop type. The regression coefficient between field measured α_{PAR} and plant chlorophyll for corn is 6.3, while for soybean it is 9.7. This demonstrates the differences in photosynthetic efficiency between corn and soybean. Using the MODIS-derived α_{PAR} (Fig. 4.5). I find the same relationship, and similar regression coefficients (6.2 for corn, and 9.7 for soybean), these relationships are shown in Table 4.2.

	Maize (corn)		**Soybean**	
	R^2	**Slope**	R^2	**Slope**
Field measured	0.93	6.3	0.92	9.7
MODIS-derived	0.89	6.2	0.92	9.7

Table 4.2: Relationship between α_{PAR} and canopy chlorophyll content by crop type

Taking the maximum α_{PAR} and the maximum plant chlorophyll content for each site and year and comparing it to the yield values (shown in Table 4.3) shows that irrigated sites have higher yields and α_{PAR} values than rainfed sites. Similarly to the yields in Chapter 3, corn yields are higher than soybean yields.

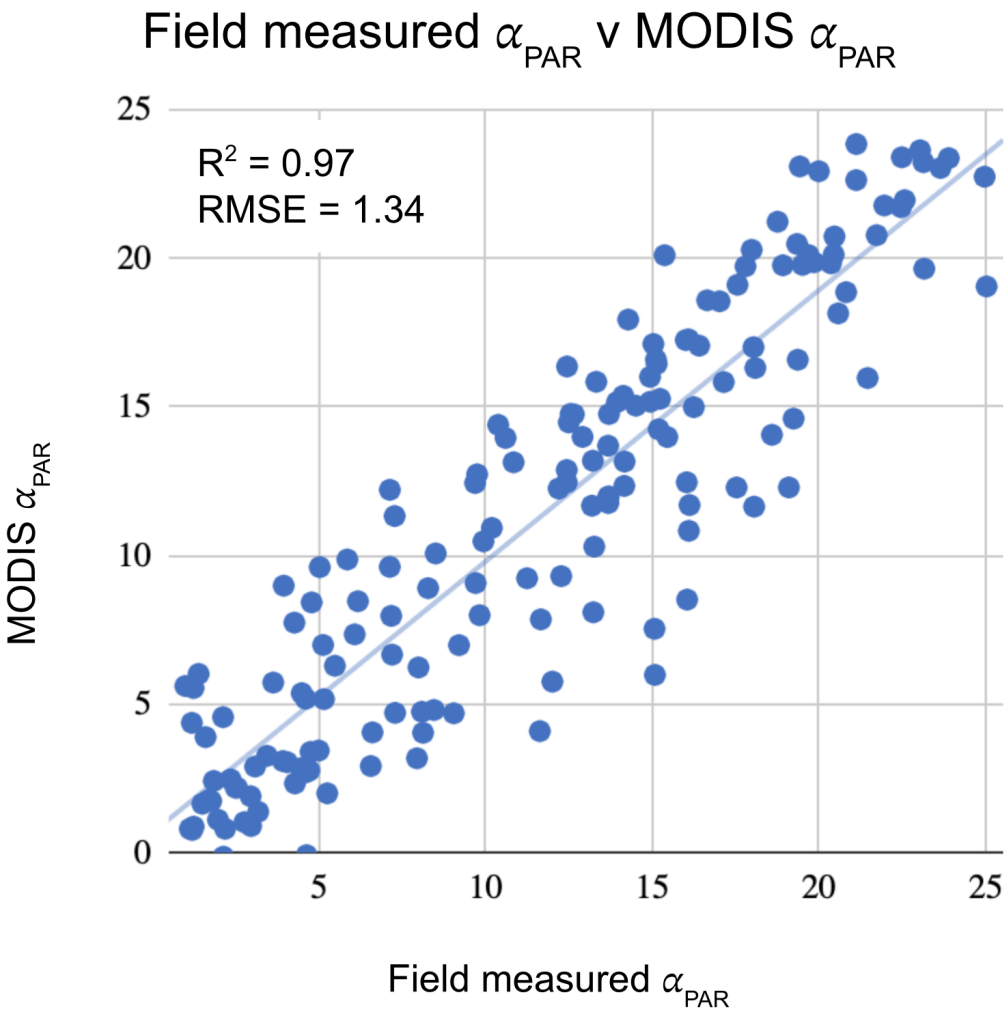

Figure 4.3: Relationship between field measured α_{PAR} and α_{PAR} derived from MODIS 500m BRDF corrected surface reflectance.

	Yield (t/ha)	**Max α_{PAR}**	**Max plant chlor. (g/m^2)**
Irrigated corn	14	23.04	3.61
Rainfed corn	7.72	17.94	2.45
Irrigated soybean	3.99	23.84	2.30
Rainfed soybean	3.32	15.27	1.80

Table 4.3: Relationships between crop type and irrigation method, maximum α_{PAR}, maximum chlorophyll content, and yield.

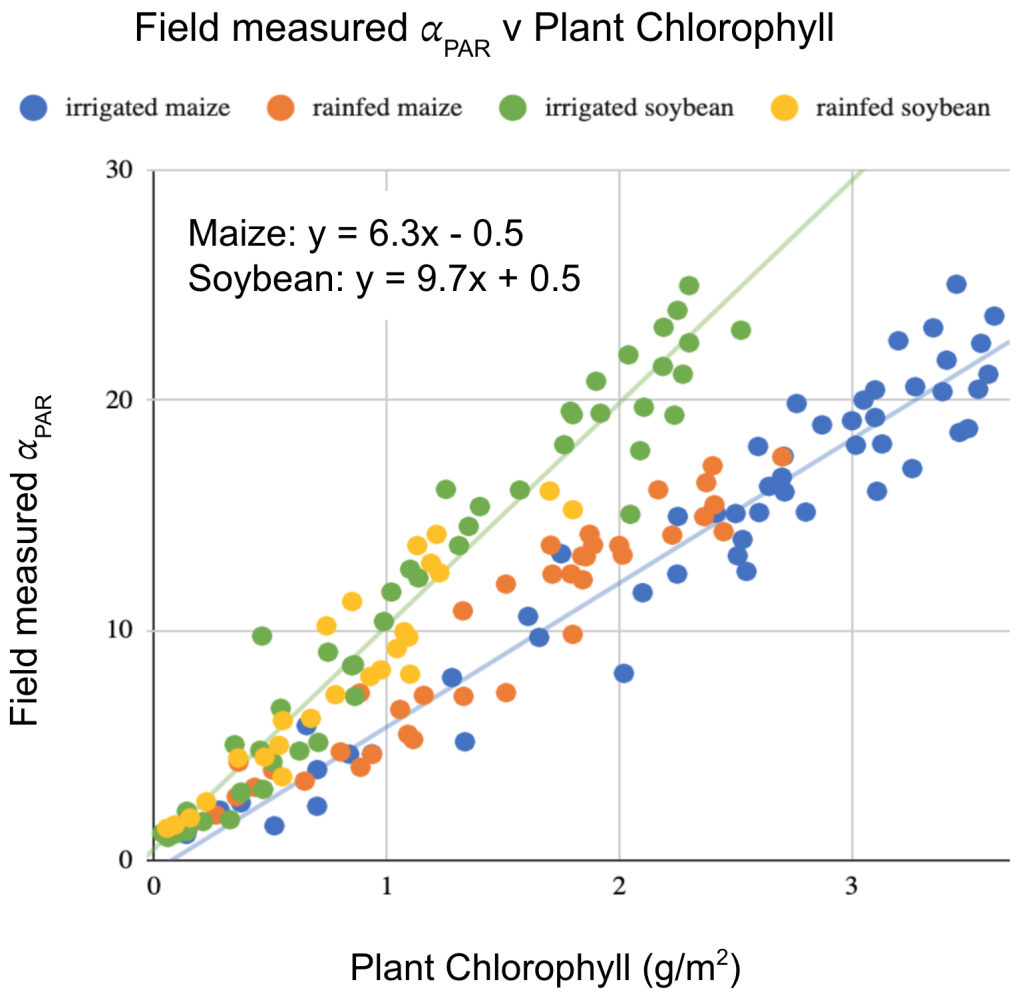

Figure 4.4: Relationship between field measured α_{PAR} and plant chlorophyll (g/m^2).

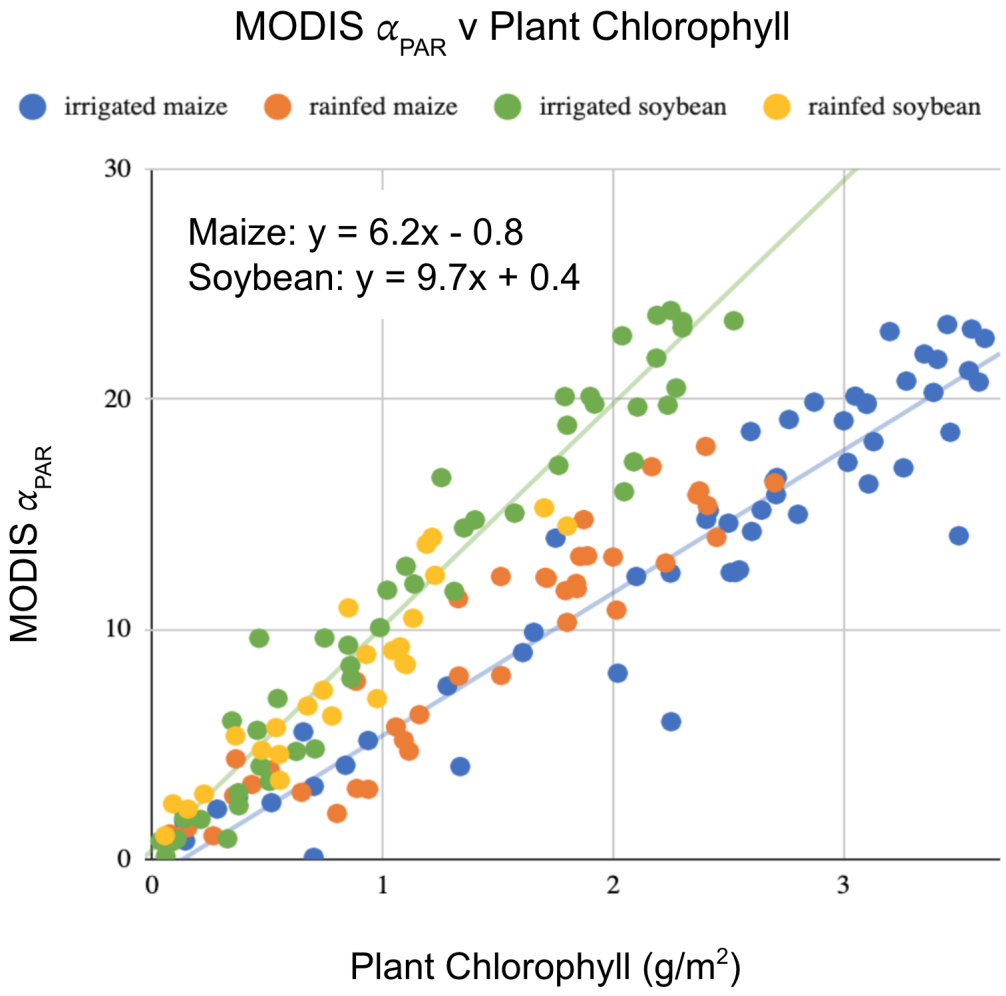

Figure 4.5: Relationship between MODIS-derived α_{PAR} and plant chlorophyll categorized by crop type and irrigation condition.

4.6.2 Crop type and condition analysis

Figure 4.6 shows the seasonal curve of MODIS-derived α_{PAR} for all data for all fields from 2002-2005. As expected, peak α_{PAR} occurs between DOY 190 and 205 for corn, with irrigated fields having higher α_{PAR} than rainfed fields, and peak α_{PAR} occurs between DOY 220 and 240 for soybean, again with irrigated fields having higher values than rainfed fields. Figure 4.7 shows the MODIS-derived α_{PAR} over eastern Nebraska for DOY 201 in 2003 (20 July 2003), which is near when we would see peak α_{PAR} in all three fields as shown in Figure 4.6. High values of α_{PAR} are shown in green, while low values of α_{PAR} are shown in purple. The high values of α_{PAR} align with the locations of corn and soybean fields in the region, while the low values align with non-cropped areas.

Figure 4.8 shows the time series of two cropped grid cells in Saunders County, Nebraska, marked by the black dots in Figure 4.2. Both grid cells, located at 41°11′23.4414″ N, 96°29′35.358″ W and 41°9′52.0842″ N, 96°24′22.035″ W, are cropped on a corn-soybean rotation. Table 4.4 shows the DOY of maximum α_{PAR} and the corresponding crop mask label. Selecting DOY 215 $\pm$ 3 (beginning of August) as the threshhold between corn and soybean classification, the first location shows 81% accuracy with the CDL. For the second location, the same threshhold gives only 64% accuracy with the CDL.

As in the experimental sites, the magnitude of α_{PAR} is indistinguishable between corn and soybean, however the timing of the peak can distinguish between the two crops. The magnitude of α_{PAR} gives more information about the irriga-

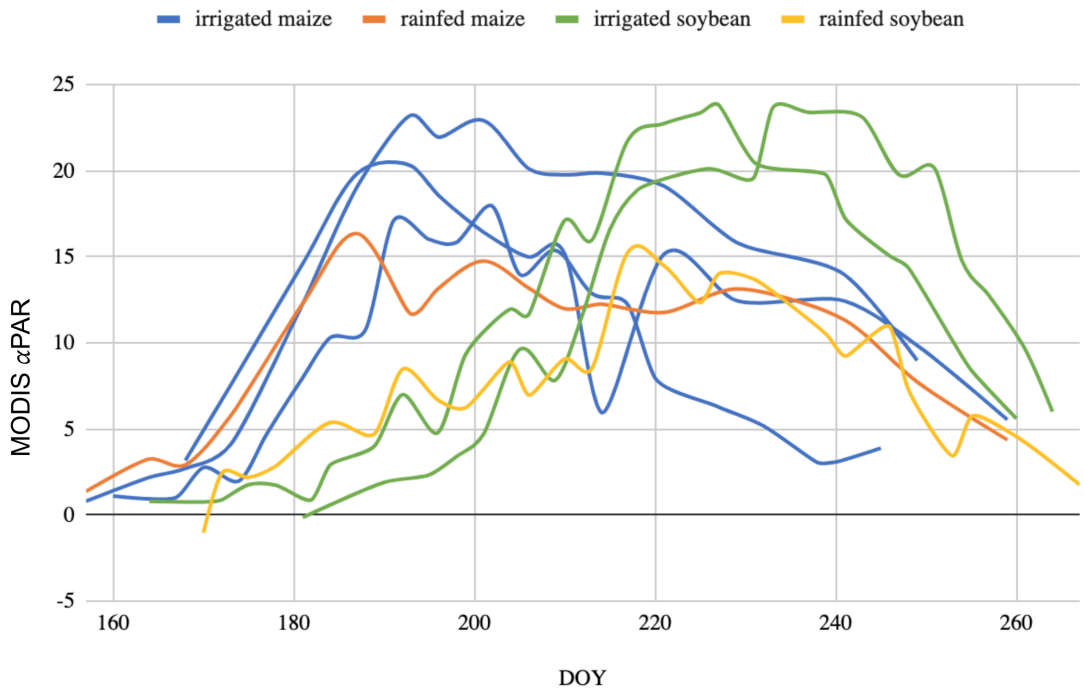

Figure 4.6: Seasonal signal of MODIS-derived α_{PAR} for corn and soybean at the three sites in Nebraska for the years 2002-2005, smoothed for display purposes.

tion and management conditions. Based on the size and shape of the fields that each grid cell belongs to, I infer that both grid cells belong to rainfed fields, and therefore the spikes in α_{PAR} are due to excessive rain or drought, extreme temperatures, and/or other management practices. According to precipitation data from the National Oceanic and Atmospheric Administration National Centers for Environmental Information (NOAA NCEI), shown in Figure 4.9, peaks in α_{PAR} in 2010, and from 2014 to 2019 may correspond to above average precipitation, however spikes in other years do not correspond to the average observed precipitation over east central Nebraska.

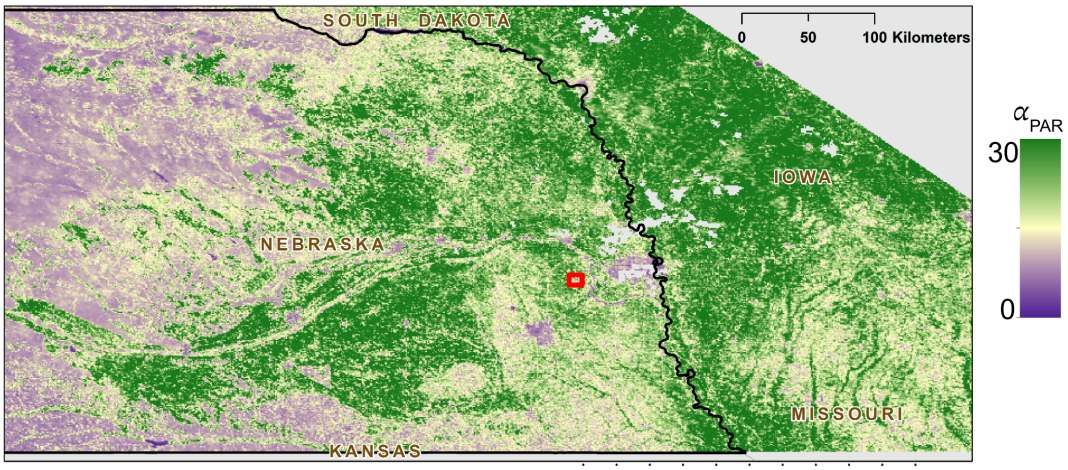

Figure 4.7: MODIS α_{PAR} calculated for DOY 201 (July 20th) in 2003 over eastern Nebraska where the majority of corn and soybean fields are. The location of the 3 sites is marked by the red box.

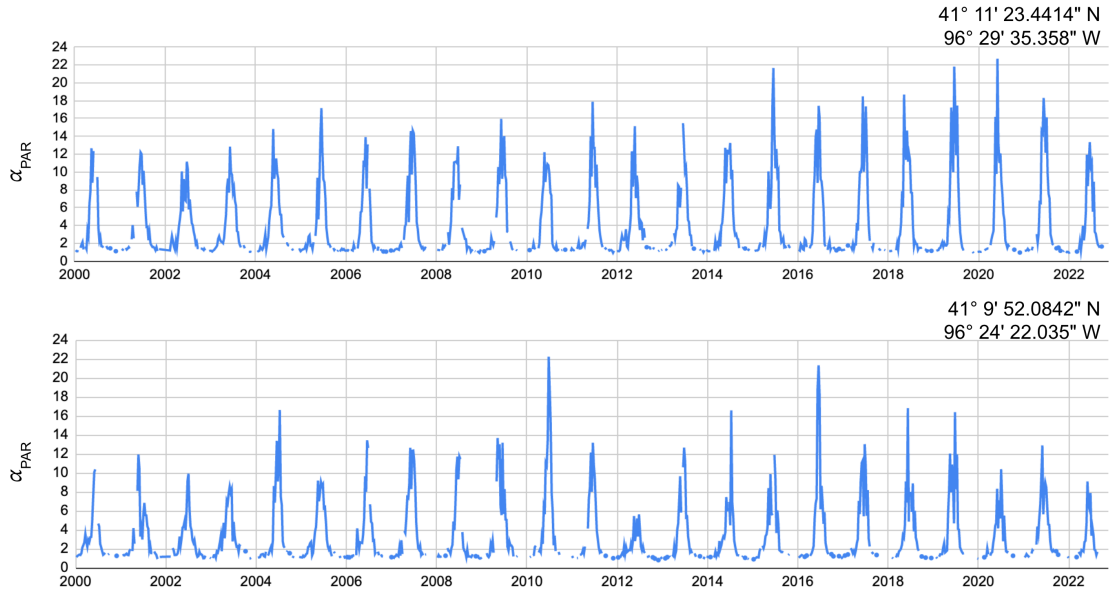

Figure 4.8: Time series of two neighboring cropped MODIS grid cells in Saunders, County, Nebraska, marked in black dots on Figure 4.2.

Year	41°11′23″ N, 96°29′35″ W Max α_{PAR}	Crop est.	CDL	41°9′52″ N, 96°24′22″ W Max α_{PAR}	Crop est.	CDL
2000	209	corn	corn	201	corn	corn
2001	193	corn	corn	217	unknown	soybean
2002	233	soybean	soybean	225	soybean	soybean
2003	201	corn	corn	209	corn	corn
2004	241	soybean	soybean	193	corn	corn
2005	185	corn	corn	217	unknown	soybean
2006	217	unknown	soybean	209	corn	corn
2007	201	corn	corn	225	soybean	soybean
2008	217	unknown	corn	225	soybean	soybean
2009	177	corn	corn	209	corn	corn
2010	225	soybean	soybean	193	corn	corn
2011	217	unknown	soybean	217	unknown	soybean
2012	225	soybean	soybean	193	corn	corn
2013	225	soybean	soybean	217	unknown	corn
2014	241	soybean	soybean	233	soybean	soybean
2015	225	soybean	soybean	217	unknown	soybean
2016	217	unknown	soybean	217	unknown	soybean
2017	225	soybean	soybean	209	corn	corn
2018	209	corn	corn	177	corn	corn
2019	225	soybean	soybean	217	unknown	soybean
2020	233	soybean	soybean	201	corn	corn
2021	201	corn	corn	209	corn	corn
2022	201	corn	corn	217	unknown	soybean

Table 4.4: Crop type identification according to phenology of α_{PAR} compared to the CDL. The threshold I chose according to the results from the experimental sites is DOY 215 ± 3, and therefore DOY 217 is always classified as unknown. Using BRDF from combined Terra and Aqua might improve my classification.

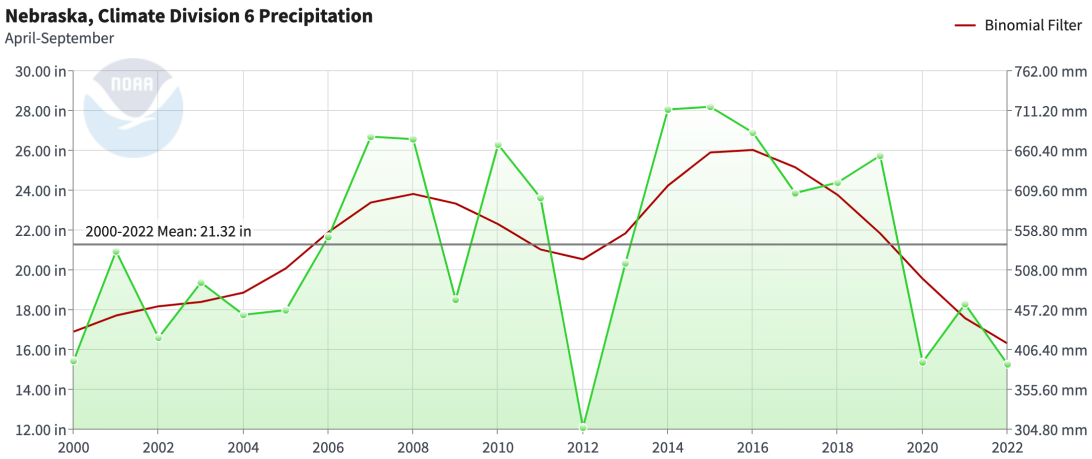

Figure 4.9: NOAA National Centers for Environmental Information (NCEI) precipitation time series from 2000 to 2022 (NCEI, 2023)

4.7 Discussion

MODIS-retrieved α_{PAR} has good correspondence to ground measured α_{PAR} ($R^2 = 0.97$, RMSE $= 11\%$), indicating that for large cropped fields and regions where the spatial resolution of the sensor is not the most limiting factor, MODIS α_{PAR} can be used in the same way that MODIS surface reflectance and vegetation indices are used for studying large cropped areas (Bolton and Friedl, 2013; Sakamoto et al., 2013; Kouadio et al., 2014). This is especially exciting because α_{PAR} has a stronger linear, and importantly crop specific, relationship to plant or canopy chlorophyll content (coefficients 6.2 and 9.7 for corn and soybean respectively), than surface reflectance or VIs alone do. It also indicates that combined MODIS-VIIRS (Visible Infrared Imaging Radiometer Suite) products (Xiong and Butler, 2020) will continue to provide value for large scale crop yield modeling studies given that Terra exited its orbital track in early 2023 and Aqua will soon follow.

The difference in magnitude of α_{PAR} between corn and soybean is negligible under comparable irrigation conditions according to the three experimental sites. However, the phenology of α_{PAR} can be used to identify corn or soybean according to the DOY of peak α_{PAR} in the exact same way as a vegetation index (e.g., NDVI) would be used (Xian et al., 2009; Wang et al., 2020b). Once the crop type has been identified according to its phenology, irrigation or soil moisture conditions can be inferred from the relative magnitudes of the signal over time.

Given the size and shape of the two cropped grid cells, I infer that they are both rainfed rather than irrigated, therefore for this analysis, I used average, regional precipitation as irrigation condition information. My findings indicate that the two sites are not in perfect correspondence with average regional rainfall, thus I suggest that comparing against soil moisture explicitly, or minimally temperature and more precise precipitation data would likely better explain the differences observed in the 22-year time series of α_{PAR} at the two cropped locations near the experimental sites.

4.8 Conclusions

Through this work, I have shown that MODIS is able to capture the same relationship between α_{PAR} and Plant Chlorophyll as the field measured α_{PAR}, for different crop types and irrigation management. MODIS-derived α_{PAR} has an R^2 of 0.97 compared to field measured α_{PAR}. I have also shown that MODIS α_{PAR} can be used to identify corn and soybean type based on the phenology of the two crops between 64% and 81% accuracy, and that the magnitude of α_{PAR} can be used to

infer changes in irrigation type, annual rainfall, extreme temperatures, or changes in management practice. However, attributing the specific causes requires more comprehensive validation information.

The implications of being able to use MODIS-derived α_{PAR} in place of field measurements for fields of this size and larger, are that α_{PAR} can be used along with PAR in the empirical crop models I developed in Chapter 3 rather than a vegetation index to estimate yields at the county scale. This is particularly because α_{PAR} has a crop specific linear relationship with plant chlorophyll content, which also has a linear relationship with crop yield (Wood et al., 1993).

Chapter 5: Conclusions

In the following sections I will discuss how my research addressed the dissertation research questions I asked in Chapter 1, the significance of my findings, and the new questions that have arisen from this work. I will also discuss some of the broader issues that we researchers should consider for the future of our work and how it can best serve human interests.

5.1 Summary of findings and significance

5.1.1 Chapter 2: Limitations and best uses of machine learning algorithms in empirical models

The machine learning based empirical model performs well compared to the physically-based LUT approach of the MODIS PAR product, especially considering that my machine learning based models do not include ancillary atmospheric information, whereas the LUT approach does explicitly include variables such as water vapor, and aerosol optical depth Wang et al. (2020a). Additionally, all of my machine learning-based models were run on a desktop computer, rather than in a high performance computing (HPC) environment. The focus of my experiments was

to see how well surface retrievals could be accomplished using an empirical model with top of atmosphere (TOA) only inputs, but future work comparing the computational performance of the entire workflow of each of these methods would lend valuable insight into the true portability of my methods.

For future development, the most practical model to use is the Bagged Tree model. The Bagged Tree model is quick to train depending on the number of trees in the bag, it has minimal hyperparameters to tune. Furthermore, of the nine models, the Bagged Tree best models the multi-modality of the ground data. The MLP and GPR capture the multi-modality, but they show higher relative errors at the lower PAR and SSR values compared to the Bagged Tree. However, more complex models, such as convolutional neural networks, deep learning networks, mixture density networks, and others are constantly being improved upon (Yuan et al., 2020). Some are being made more transparent, while others have been optimized to reduce the computational power required (Al-Jarrah et al., 2015). Although some of the methods are better suited for classification, pattern or object recognition than regression (Bishop, 2006), it would still be prudent to experiment with these methods in the future, as they are held in high regard for their potential solving our most challenging problems. Care must be taken, however, to be sure to understand the limitations of these methods when interpreting their results. It may be tempting to over train and over tune a model to get "good" results, but we researchers must always be able to explain the physical mechanisms that are being modeled with these statistical methods. It is important to avoid reporting spurious results and perpetuating the idea that machine learning is a "magic bullet" when it comes to

the complex problems of the Earth Science.

5.1.2 Chapter 3: Benefits and caveats of explicitly adding PAR to empirical yield models

By adding PAR explicitly into an empirical crop yield model, R^2_{adj} increases from approximately 0.60 to between 0.81 and 0.80 for corn and soy, respectively. The results of county-level yield modeling align with field-level studies (Johnson, 2016; Franch et al., 2019; Skakun et al., 2021). Specifically, for both corn and soy, the VI-only model shows a maximum R^2_{adj} of around 0.60, which corresponds very well to (Johnson, 2016)'s findings of the correlation of VI to yield, even at the coarse MODIS scale.

The addition of PAR to the model improves on a SR or VI based models, even at MODIS scale and in the United States where radiation is not the limiting factor for vegetation growth (Nemani et al., 2003; Milesi et al., 2005; Running et al., 2004). Given this improvement, in the United States, where agricultural systems are highly managed and radiation is not the limiting factor, my findings indicate that the addition of PAR to empirical yield models in regions in the world where vegetation is radiation limited may add significant information. However, a limiting factor to that work is the spatial scale of PAR data. The current MODIS product (MCD18A2) is limited to 1 km spatial resolution due to the ancillary atmospheric and reanalysis input data that the algorithm requires.

Future work should make use of finer spatial resolution satellite systems so

that it can be directed at the field scale, and eventually the sub-field scale. These levels are often of greater interested to farmers and the government entities responsible for overseeing and managing agricultural polices and practices. This will require either using limited finer resolution observations to resolve the MODIS product (MCD18A2) down to a finer spatial resolution, or generating satellite derived estimates of PAR from finer resolution satellite observations. A variety of different methods are available for both options, however each will require additional validation before they can be incorporated into empirical crop models.

Additionally, as in Chapter 2, more sophisticated machine learning algorithms could be used in place of linear regression. This may provide better resolution of non-linearities in the relationships between indicators and crop yield, particularly for crop types that are less well modeled using linear regression. However, machine learning algorithms require large amounts of labeled training data which is not often available, especially over long time periods and at enough spatially distinct locations.

5.1.3 Chapter 4: Potential and caveats for large scale crop modeling using coarse scale satellite-derived absorption coefficient of PAR

MODIS-derived αPAR matches field measured αPAR extremely well for these three large fields from Chapter 4, with an R^2 of 0.97. MODIS-derived α_{PAR} also shows the same crop specific relationships with canopy chlorophyll content (regression coefficient of 6.2 and 9.7 for corn and soybean respectively). The phenology of

α_{PAR} is also a good indicator of crop type, with an accuracy between 64% and 81% for two randomly selected locations. Finally, α_{PAR} can be used to infer changes in crop conditions, including irrigation type, average rainfall, excessive temperatures, or other management practices. This finding is important for future work incorporating αPAR into empirical crop yield models, such as those I used in Chapter 3, and analyzing any long term global trends in the absorption of PAR by different crop species. It also sets the precedent for the use of combined MODIS-VIIRS data moving forward as the MODIS sensors aboard Terra and Aqua will soon be impractical to continue using.

Further work should take a two-pronged approach. 1) Determine the degree to which Harmonized Landsat Sentinel (HLS) α_{PAR} matches in field measurements for further work with smaller fields and to distinguish between the management practices in neighboring fields. 2) Expand the work to learn the crop specific relationships between α_{PAR} and canopy chlorophyll content for other major crops that make up the global food supply, and determine the most appropriate data to use to capture information about soil moisture, soil nutrients, and other major management practices used for cultivated crops in order to improve yield estimates.

5.2 Looking toward the future

5.2.1 Implications for data and research autonomy

One opportunistic, yet important, aspect of this research is that none of it required high performance computing (HPC) systems or graphical processing units

(GPUs). All of the modeling and analysis was done on a desktop or laptop computer with a standard central processing unit (CPU). This means that all of this research could be recreated by researchers outside of large, well-funded universities and national laboratories. Researchers anywhere in the world, with their own regional specific expertise could use these methods for their own work. I tested all of these methods using MODIS data due to its widespread availability and the scientific quality of the data, but it is not an inherent requirement that MODIS data be used. And in fact, since MODIS is at the end of its lifetime, recreating this research with other available satellite observations and ground data from other parts of the world is an important next step.

Additionally, the empirical crop models I used in Chapter 3 and the crop specific relationships between α_{PAR} and canopy chlorophyll content from Chapter 4, could be tuned/explored for other crops, including wheat, rice, sorghum, millet, and others that represent significant portions of the global food supply. The models from Chapter 3 can be used for year to year comparisons and trend analysis, or for near real time (NRT) monitoring, which, if implemented by regional stakeholders and the science teams that inform policy makers, could be used in conjunction with other early warning and monitoring efforts to help state and local governments deal with potential food shortages or surpluses.

5.2.2 Crop modeling with PAR at very high resolutions

With the rise in commercially available Very High Resolution (VHR) satellites, hopefully scientists and the private sector can join together in making these data available for scientific use, as well as the private sector opening themselves up for creating and supporting instruments that can meet the scientific standards of instruments like MODIS, VIIRS, Landsat, and Sentinel.

If higher resolution PAR products were available, in PAR limited regions, in regions which enact clean air policies/measures, or in response to human efforts to combat climate change through geoengineering Lockley et al. (2020), the impact of global dimming and brightening Wild et al. (2013), and subsequently changes in PAR, on crop yields could be better studied and monitored.

5.2.3 Climate change impacts on food security

This work was all based in the United States because of the richness of available data. However, there are many regions of the world, such as Northern Africa and the Middle East, sub Sahelian West Africa, and East Africa (Nakalembe; Kerner et al., 2020; Nakalembe et al., 2019), where organizations such as NASA Harvest and the Famine Early Warning System Network (FEWSNET) have explicit interest in order to support their missions around global food security. Obtaining data over these regions presents an inconsequential challenge (due to cloud cover, historic exclusion from the data record, etc.), but with the expansion of commercial satellite data and cooperative agreements between companies and academia or the U.S. federal

government, these may be used more widely for the benefit of society.

Even for the United States and other data rich regions that export major commodities globally, organizations with national security interests are also users and stakeholders in this work. Especially as climate change continues to impact global temperatures and rainfall (Tirado et al., 2010; Barnett, 2011; Misra, 2014), and as countries begin implementing geoengineering strategies to mitigate the effects of climate change (Liu et al., 2021; Lockley et al., 2020), the consequences, intended and otherwise, of such efforts must be studied.

Bibliography

SA Ackerman and R Frey. MODIS atmosphere L2 cloud mask product (35_L2). *NASA MODIS Adaptive Processing System, NASA Goddard Space Flight Center: Greenbelt, MD, USA*, 2015.

Omar Y Al-Jarrah, Paul D Yoo, Sami Muhaidat, George K Karagiannidis, and Kamal Taha. Efficient machine learning for big data: A review. *Big Data Research*, 2(3):87–93, 2015.

PB Alton, R Ellis, SO Los, and PR North. Improved global simulations of gross primary product based on a separate and explicit treatment of diffuse and direct sunlight. *Journal of Geophysical Research: Atmospheres*, 112(D7), 2007.

Mv C Anderson. Radiation and crop structure. *Plant photosynthetic production. Manual of methods.*, pages 412–466, 1971.

John A Augustine, Gary B Hodges, Christopher R Cornwall, Joseph J Michalsky, and Carlos I Medina. An update on SURFRAD: The GCOS surface radiation budget network for the continental United States. *Journal of Atmospheric and Oceanic Technology*, 22(10):1460–1472, 2005.

Dianna K Bagnall, John F Shanahan, Archie Flanders, Cristine LS Morgan, and C Wayne Honeycutt. Soil health considerations for global food security. *Agronomy Journal*, 113(6):4581–4589, 2021.

Jon Barnett. Dangerous climate change in the pacific islands: food production and food security. *Regional Environmental Change*, 11:229–237, 2011.

Bruno Basso, Davide Cammarano, and Elisabetta Carfagna. Review of crop yield forecasting methods and early warning systems. In *Proceedings of the first meeting of the scientific advisory committee of the global strategy to improve agricultural and rural statistics, FAO Headquarters, Rome, Italy*, volume 18, page 19, 2013.

Katja Berger, Clement Atzberger, Martin Danner, Guido D'Urso, Wolfram Mauser, Francesco Vuolo, and Tobias Hank. Evaluation of the prosail model capabilities for future hyperspectral model environments: A review study. *Remote Sensing*, 10(1):85, 2018.

Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

Douglas K Bolton and Mark A Friedl. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology*, 173:74–84, 2013.

JS Boyer. Differing sensitivity of photosynthesis to low leaf water potentials in corn and soybean. *Plant physiology*, 46(2):236–239, 1970.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

François-Marie Bréon and Eric Vermote. Correction of modis surface reflectance time series for brdf effects. *Remote Sensing of Environment*, 125:1–9, 2012.

Meredith GL Brown, Sergii Skakun, Tao He, and Shunlin Liang. Intercomparison of Machine-Learning methods for estimating Surface Shortwave and Photosynthetically Active Radiation. *Remote Sensing*, 12(3):372, 2020.

Manuel L Campagnolo and Enrique L Montano. Estimation of effective resolution for daily MODIS gridded surface reflectance products. *IEEE Transactions on Geoscience and Remote Sensing*, 52(9):5622–5632, 2014.

James B Campbell and Randolph H Wynne. *Introduction to remote sensing*. Guilford Press, 2011.

Gustavo Camps-Valls, Luis Gómez-Chova, Jordi Muñoz-Marí, Joan Vila-Francés, Julia Amorós-López, and Javier Calpe-Maravilla. Retrieval of oceanic chlorophyll concentration with relevance vector machines. *Remote Sensing of Environment*, 105(1):23–33, 2006.

Corinne Carter and Shunlin Liang. Evaluation of ten machine learning methods for estimating terrestrial evapotranspiration from remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 78:86–92, 2019.

Subrahmanyan Chandrasekhar. *Radiative Transfer*. Courier Corporation, 1960.

Susan J Cheng, Gil Bohrer, Allison L Steiner, David Y Hollinger, Andrew Suyker, Richard P Phillips, and Knute J Nadelhoffer. Variations in the influence of diffuse light on gross primary productivity in temperate ecosystems. *Agricultural and Forest Meteorology*, 201:98–110, 2015.

Martin Claverie, Junchang Ju, Jeffrey G Masek, Jennifer L Dungan, Eric F Vermote, Jean-Claude Roger, Sergii V Skakun, and Christopher Justice. The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote Sensing of Environment*, 219:145–161, 2018.

James R Ehleringer and Thure E Cerling. C3 and C4 photosynthesis. *Encyclopedia of global environmental change*, 2(4):186–190, 2002.

J-B Féret, AA Gitelson, SD Noble, and S Jacquemoud. Prospect-d: Towards modeling leaf optical properties through a complete lifecycle. *Remote Sensing of Environment*, 193:204–215, 2017.

Jean-Baptiste Féret, Katja Berger, Florian De Boissieu, and Zbyněk Malenovskỳ. Prospect-pro for estimating content of nitrogen-containing leaf proteins and other carbon-based constituents. *Remote Sensing of Environment*, 252:112173, 2021.

Yolanda. M. Fernandez-Ordoñez and J. Soria-Ruíz. Maize crop yield estimation with remote sensing and empirical models. *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3035–3038, 2017. doi: 10.1109/IGARSS.2017.8127638.

Belen Franch, Eric F Vermote, Sergii Skakun, Jean-Claude Roger, Inbal Becker-Reshef, Emilie Murphy, and C Justice. Remote sensing based yield monitoring: Application to winter wheat in United States and Ukraine. *International Journal of Applied Earth Observation and Geoinformation*, 76:112–127, 2019.

Feng Gao, Martha Anderson, Craig Daughtry, and David Johnson. Assessing the variability of corn and soybean yields in central Iowa using high spatiotemporal resolution multi-satellite imagery. *Remote Sensing*, 10(9):1489, 2018.

Anatoly Gitelson, Andrés Viña, Alexei Solovchenko, Timothy Arkebauer, and Yoshio Inoue. Derivation of canopy light absorption coefficient from reflectance spectra. *Remote Sensing of Environment*, 231:111276, 2019.

Anatoly Gitelson, Timothy Arkebauer, Andrés Viña, Sergii Skakun, and Yoshio Inoue. Evaluating plant photosynthetic traits via absorption coefficient in the photosynthetically active radiation region. *Remote Sensing of Environment*, 258: 112401, 2021.

Anatoly A Gitelson. Wide dynamic range vegetation index for remote quantification of biophysical characteristics of vegetation. *Journal of plant physiology*, 161(2): 165–173, 2004.

Anatoly A Gitelson and John A Gamon. The need for a common basis for defining light-use efficiency: Implications for productivity estimation. *Remote Sensing of Environment*, 156:196–201, 2015.

Anatoly A Gitelson, Yi Peng, Timothy J Arkebauer, and James Schepers. Relationships between gross primary production, green LAI, and canopy chlorophyll content in maize: Implications for remote sensing of primary production. *Remote Sensing of Environment*, 144:65–72, 2014.

Anatoly A Gitelson, Yi Peng, Timothy J Arkebauer, and Andrew E Suyker. Productivity, absorbed photosynthetically active radiation, and light use efficiency in crops: Implications for remote sensing of crop primary production. *Journal of Plant Physiology*, 177:100–109, 2015.

Samuel N Goward, Compton J Tucker, and Dennis G Dye. North american vegetation patterns observed with the NOAA-7 advanced very high resolution radiometer. *Vegetatio*, 64(1):3–14, 1985.

Lianhong Gu, Dennis D Baldocchi, Steve C Wofsy, J William Munger, Joseph J Michalsky, Shawn P Urbanski, and Thomas A Boden. Response of a deciduous forest to the mount pinatubo eruption: Enhanced photosynthesis. *Science*, 299 (5615):2035–2038, 2003.

Sheng Gui, Shunlin Liang, Kaicun Wang, Lin Li, and Xiaotong Zhang. Assessment of three satellite-estimated land surface downwelling shortwave irradiance data sets. *IEEE Geoscience and Remote Sensing Letters*, 7(4):776–780, 2010.

Weiguo Han, Zhengwei Yang, Liping Di, and Richard Mueller. Cropscape: A web service based application for exploring and disseminating us conterminous geospatial cropland data products for decision support. *Computers and Electronics in Agriculture*, 84:111–123, 2012.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

A Huete, C Justice, and H Liu. Development of vegetation and soil indices for MODIS-EOS. *Remote Sensing of Environment*, 49(3):224–234, 1994.

Alfredo Huete, Kamel Didan, Tomoaki Miura, E Patricia Rodriguez, Xiang Gao, and Laerte G Ferreira. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, 83(1-2):195–213, 2002.

AR Huete, HQ Liu, KV Batchily, and WJDA Van Leeuwen. A comparison of vegetation indices over a global set of TM images for EOS-MODIS. *Remote Sensing of Environment*, 59(3):440–451, 1997.

Peter J Irvine, Ben Kravitz, Mark G Lawrence, and Helene Muri. An overview of the Earth system science of solar geoengineering. *Wiley Interdisciplinary Reviews: Climate Change*, 7(6):815–833, 2016.

Stéphane Jacquemoud, Frédéric Baret, and JF Hanocq. Modeling spectral and bidirectional soil reflectance. *Remote sensing of Environment*, 41(2-3):123–132, 1992.

Stéphane Jacquemoud, Wout Verhoef, Frédéric Baret, Cédric Bacour, Pablo J Zarco-Tejada, Gregory P Asner, Christophe François, and Susan L Ustin. Prospect+ sail models: A review of use for vegetation characterization. *Remote sensing of environment*, 113:S56–S66, 2009.

David M Johnson. An assessment of pre-and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sensing of Environment*, 141:116–128, 2014.

David M Johnson. A comprehensive assessment of the correlations between field crop yields and commonly used MODIS products. *International Journal of Applied Earth Observation and Geoinformation*, 52:65–81, 2016.

Christopher O Justice, JRG Townshend, BN Holben, and et CJ Tucker. Analysis of the phenology of global vegetation using meteorological satellite data. *International Journal of Remote Sensing*, 6(8):1271–1318, 1985.

Christopher O Justice, Miguel O Román, Ivan Csiszar, Eric F Vermote, Robert E Wolfe, Simon J Hook, Mark Friedl, Zhuosen Wang, Crystal B Schaaf, Tomoaki Miura, et al. Land and cryosphere products from Suomi NPP VIIRS: Overview and status. *Journal of Geophysical Research: Atmospheres*, 118(17):9753–9765, 2013.

Kasturi Devi Kanniah, Jason Beringer, Peter North, and Lindsay Hutley. Control of atmospheric particles on diffuse radiation and terrestrial plant productivity: A review. *Progress in Physical Geography*, 36(2):209–237, 2012.

Leonid Katkovsky, Anton Martinov, Volha Siliuk, Dimitry Ivanov, and Alexander Kokhanovsky. Fast atmospheric correction method for hyperspectral data. *Remote Sensing*, 10(11):1698, 2018.

Hannah Kerner, Catherine Nakalembe, and Inbal Becker-Reshef. Field-level crop type classification with k Nearest Neighbors: A baseline for a new Kenya smallholder dataset. *arXiv preprint arXiv:2004.03023*, 2020.

Kerrey Kerr-Enskat. U.S. soy exports earn record $40.42 billion on second-highest volumes of 71.79 MMT in marketing year 21/22, Dec 2022. URL `https://ussec.org/`.

Louis Kouadio, Nathaniel K Newlands, Andrew Davidson, Yinsuo Zhang, and Aston Chipanshi. Assessing the performance of modis ndvi and evi for seasonal crop yield forecasting at the ecodistrict scale. *Remote Sensing*, 6(10):10193–10214, 2014.

Nataliia Kussul, Mykola Lavreniuk, Sergii Skakun, and Andrii Shelestov. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14(5):778–782, 2017.

Patrice Latinne, Olivier Debeir, and Christine Decaestecker. Limiting the number of trees in random forests. In *International workshop on multiple classifier systems*, pages 178–187. Springer, 2001.

Miguel Lázaro-Gredilla and Michalis K Titsias. Variational Heteroscedastic Gaussian Process Regression. In *ICML*, pages 841–848, 2011.

Miguel Lázaro-Gredilla, Michalis K Titsias, Jochem Verrelst, and Gustavo Camps-Valls. Retrieval of biophysical parameters with heteroscedastic Gaussian processes. *IEEE Geoscience and Remote Sensing Letters*, 11(4):838–842, 2014.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

Mireille Lefèvre, Armel Oumbe, Philippe Blanc, Bella Espinar, Benoît Gschwind, Zhipeng Qu, Lucien Wald, Marion Schroedter Homscheidt, Carsten Hoyer-Klick, Antti Arola, et al. Mcclear: A new model estimating downwelling solar radiation at ground level in clear-sky conditions. *Atmospheric Measurement Techniques*, 6: 2403–2418, 2013.

Shunlin Liang. *Quantitative remote sensing of land surfaces*, volume 30. John Wiley & Sons, 2005.

Shunlin Liang, Tao Zheng, Ronggao Liu, Hongliang Fang, Si-Chee Tsay, and Steven Running. Estimation of incident photosynthetically active radiation from Moderate Resolution Imaging Spectrometer data. *Journal of Geophysical Research: Atmospheres*, 111(D15), 2006.

Zhaochen Liu, Xianmei Lang, and Dabang Jiang. Impact of stratospheric aerosol injection geoengineering on the summer climate over East Asia. *Journal of Geophysical Research: Atmospheres*, 126(22):e2021JD035049, 2021.

David Lobell, K Cahill, Christopher Field, et al. Weather-based yield forecasts developed for 12 California crops. *California Agriculture*, 60(4):211–215, 2006.

Andrew Lockley, Doug MacMartin, and Hugh Hunt. An update on engineering issues concerning stratospheric aerosol injection for geoengineering. *Environmental Research Communications*, 2(8):082001, 2020.

Michael Marshall, Kevin Tu, and Jesslyn Brown. Optimizing a remote sensing production efficiency model for macro-scale GPP and yield estimation in agroecosystems. *Remote Sensing of Environment*, 217:258–271, 2018.

Jordane A Mathieu and Filipe Aires. Assessment of the agro-climatic indices to improve crop yield forecasting. *Agricultural and forest meteorology*, 253:15–30, 2018.

Sonal Mathur, Divya Agrawal, and Anjana Jajoo. Photosynthesis: response to high temperature stress. *Journal of Photochemistry and Photobiology B: Biology*, 137: 116–126, 2014.

BE Medlyn, Erwin Dreyer, D Ellsworth, M Forstreuter, PC Harley, MUF Kirschbaum, Xavier Le Roux, Pierre Montpied, J Strassemeyer, A Walcroft, et al. Temperature response of parameters of a biochemically based model of photosynthesis. II. A review of experimental data. *Plant, Cell & Environment*, 25(9): 1167–1179, 2002.

Lina M Mercado, Nicolas Bellouin, Stephen Sitch, Olivier Boucher, Chris Huntingford, Martin Wild, and Peter M Cox. Impact of changes in diffuse radiation on the global land carbon sink. *Nature*, 458(7241):1014–1017, 2009.

Cristina Milesi, Hirofumi Hashimoto, Steven W Running, and Ramakrishna R Nemani. Climate variability, vegetation productivity and people at risk. *Global and Planetary Change*, 47(2):221–231, 2005.

Anil Kumar Misra. Climate change and challenges of water and food security. *International Journal of Sustainable Built Environment*, 3(1):153–165, 2014.

MODIS Science Data Support Team. MOD021KM MODIS/Terra calibrated radiances 5-min L1B Swath 1km, a. URL `http://modaps.nascom.nasa.gov/services/about/products/c6/MOD021KM.html`.

MODIS Science Data Support Team. MOD03 MODIS/Terra geolocation fields 5-min L1A Swath 1km, b. URL `http://modaps.nascom.nasa.gov/services/about/products/c6/MOD03.html`.

MODIS Science Data Support Team. MYD021KM MODIS/Aqua calibrated radiances 5-min L1B Swath 1km, c. URL `http://modaps.nascom.nasa.gov/services/about/products/c6/MYD021KM.html`.

MODIS Science Data Support Team. MYD03 MODIS/Aqua geolocation fields 5-min L1A Swath 1km, d. URL `http://modaps.nascom.nasa.gov/services/about/products/c6/MYD03.html`.

JL Monteith. Solar radiation and productivity in tropical ecosystems. *Journal of Applied Ecology*, 9(3):747–766, 1972.

John Lennox Monteith. Climate and the efficiency of crop production in Britain. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 281(980):277–294, 1977.

David J Mulla. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems engineering*, 114(4):358–371, 2013.

Lars J Munkholm, Richard J Heck, and Bill Deen. Long-term rotation and tillage effects on soil structure and crop yield. *Soil and Tillage Research*, 127:85–91, 2013.

Ranga B Myneni, S Hoffman, Yuri Knyazikhin, JL Privette, J Glassy, Yuhong Tian, Y Wang, X Song, Y Zhang, GR Smith, et al. Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. *Remote Sensing of Environment*, 83(1-2):214–231, 2002.

Catherine Nakalembe. Application of Geospatial Science in Development Planning: The need for Geospatial Data and Information for Karamoja.

Catherine Nakalembe, Inbal Becker-Reshef, Rogerio Bonifacio, Guangxiao Hu, Michael Laurence Humber, Christina Jade Justice, John Keniston, Kenneth Mwangi, Felix Rembold, Shraddhanand Shukla, et al. A review of satellite-based global agricultural monitoring systems available for Africa. *Global Food Security*, 29:100543, 2021.

Catherine Lilian Nakalembe, Inbal Becker-Reshef, Jan Dempewolf, Lilian W Ndungu, and Kenneth Mwangi. Leveraging Earth Observations in agriculture monitoring. Building a sustained capacity development model based on Remote Sensing in Eastern Africa. *AGUFM*, 2019:PA22A–06, 2019.

NOAA NCEI. Noaa national centers for environmental information, climate at a glance: Divisional time series, May 2023. URL `https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/`.

Ramakrishna R Nemani, Charles D Keeling, Hirofumi Hashimoto, William M Jolly, Stephen C Piper, Compton J Tucker, Ranga B Myneni, and Steven W Running. Climate-driven increases in global terrestrial net primary production from 1982 to 1999. *science*, 300(5625):1560–1563, 2003.

Kenta Obata, Tomoaki Miura, Hiroki Yoshioka, Alfredo R Huete, and Marco Vargas. Spectral cross-calibration of VIIRS enhanced vegetation index with MODIS: A case study using year-long global data. *Remote Sensing*, 8(1):34, 2016.

Thais Oshiro, Pedro Perez, and José Baranauskas. How many trees in a random forest? volume 7376, 07 2012. doi: 10.1007/978-3-642-31537-4_13.

SJ Park, CS Hwang, and PLG Vlek. Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions. *Agricultural Systems*, 85(1):59–81, 2005.

Yi Peng, Anthony Nguy-Robertson, Timothy Arkebauer, and Anatoly A Gitelson. Assessment of canopy chlorophyll content retrieval in maize and soybean: Implications of hysteresis on the development of generic algorithms. *Remote Sensing*, 9(3):226, 2017.

Petra Perner. *Machine learning and data mining in pattern recognition. 8th international conference, MLDM 2012, Berlin, Germany, July 13–20, 2012. Proceedings*, volume 7376. 01 2012. doi: 10.1007/978-3-642-31537-4.

Bernard Pinty, Nadine Gobron, Jean-Luc Widlowski, Sigfried AW Gerstl, Michel M Verstraete, Mauro Antunes, Cédric Bacour, Ferran Gascon, Jean-Philippe Gastellu, Narendra Goel, et al. Radiation transfer model intercomparison (rami) exercise. *Journal of Geophysical Research: Atmospheres*, 106(D11):11937–11956, 2001.

Bernard Pinty, J-L Widlowski, Malcolm Taberner, Nadine Gobron, Michel M Verstraete, M Disney, Ferran Gascon, J-P Gastellu, L Jiang, A Kuusk, et al. Radiation transfer model intercomparison (rami) exercise: Results from the second phase. *Journal of Geophysical Research: Atmospheres*, 109(D6), 2004.

Anup K Prasad, Lim Chai, Ramesh P Singh, and Menas Kafatos. Crop yield estimation model for Iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation*, 8(1):26–33, 2006.

A Rap, DV Spracklen, L Mercado, CL Reddington, JM Haywood, RJ Ellis, OL Phillips, P Artaxo, Damien Bonal, Natalia Restrepo Coupe, et al. Fires increase amazon forest productivity through increases in diffuse radiation. *Geophysical Research Letters*, 42(11):4654–4662, 2015.

A Rap, CE Scott, CL Reddington, L Mercado, RJ Ellis, S Garraway, MJ Evans, DJ Beerling, AR MacKenzie, CN Hewitt, et al. Enhanced global primary production by biogenic aerosol via diffuse radiation fertilization. *Nature Geoscience*, 11 (9):640–644, 2018.

Matthew Clark Reeves, M Zhao, and Steven W Running. Usefulness and limits of MODIS GPP for estimating wheat yield. *International Journal of Remote Sensing*, 26(7):1403–1421, 2005.

Michael L Roderick, Graham D Farquhar, Sandra L Berry, and Ian R Noble. On the direct effect of clouds and atmospheric particles on the productivity and structure of vegetation. *Oecologia*, 129:21–30, 2001.

Steven W Running, Ramakrishna R Nemani, Faith Ann Heinsch, Maosheng Zhao, Matt Reeves, and Hirofumi Hashimoto. A continuous satellite-derived measure of global terrestrial primary production. *AIBS Bulletin*, 54(6):547–560, 2004.

Toshihiro Sakamoto, Anatoly A Gitelson, and Timothy J Arkebauer. Modis-based corn grain yield estimation model incorporating crop phenology information. *Remote Sensing of Environment*, 131:215–231, 2013.

Fadil Santosa and William W Symes. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, 1986.

Crystal B Schaaf, Feng Gao, Alan H Strahler, Wolfgang Lucht, Xiaowen Li, Trevor Tsang, Nicholas C Strugnell, Xiaoyang Zhang, Yufang Jin, Jan-Peter Muller, et al. First operational BRDF, albedo nadir reflectance products from MODIS. *Remote Sensing of Environment*, 83(1-2):135–148, 2002.

Raphael Shirley, Edward Pope, Myles Bartlett, Seb Oliver, Novi Quadrianto, Peter Hurley, Steven Duivenvoorden, Phil Rooney, Adam B Barrett, Chris Kent, et al. An empirical, Bayesian approach to modelling crop yield: Maize in USA. *Environmental Research Communications*, 2(2):025002, 2020.

Sergii Skakun, Christopher O Justice, Eric Vermote, and Jean-Claude Roger. Transitioning from MODIS to VIIRS: an analysis of inter-consistency of NDVI data sets for agricultural monitoring. *International journal of remote sensing*, 39(4): 971–992, 2018.

Sergii Skakun, Natacha I Kalecinski, Meredith GL Brown, David M Johnson, Eric F Vermote, Jean-Claude Roger, and Belen Franch. Assessing within-Field Corn and Soybean Yield Variability from WorldView-3, Planet, Sentinel-2, and Landsat 8 Satellite Imagery. *Remote Sensing*, 13(5):872, 2021.

Andy Suyker. AmeriFlux FLUXNET-1F US-Ne1 Mead - irrigated continuous maize site. 1 2022. doi: 10.17190/AMF/1871140.

Wenjun Tang, Jun Qin, Kun Yang, Xiaolei Niu, Min Min, and Shunlin Liang. An efficient algorithm for calculating photosynthetically active radiation with MODIS products. *Remote Sensing of Environment*, 194:146–154, 2017.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

M Cristina Tirado, R Clarke, LA Jaykus, A McQuatters-Gollop, and JM Frank. Climate change and food safety: A review. *Food Research International*, 43(7): 1745–1765, 2010.

CJ Tucker and PJ Sellers. Satellite remote sensing of primary production. *International journal of remote sensing*, 7(11):1395–1416, 1986.

U.S. GRAINS COUNCIL. Corn production and exports, Feb 2023. URL `https://grains.org/`.

USDA National Agriculture Statistics Service USDA NASS. Census of Agriculture. `http://www.nass.usda.gov/AgCensus`, 2023.

Patrick E Van Laake and G Arturo Sanchez-Azofeifa. Simplified atmospheric radiative transfer modelling for estimating incident PAR using MODIS atmosphere products. *Remote Sensing of Environment*, 91(1):98–113, 2004.

Patrick E Van Laake and G Arturo Sanchez-Azofeifa. Mapping PAR using MODIS atmosphere products. *Remote Sensing of Environment*, 94(4):554–563, 2005.

Wout Verhoef and Heike Bach. Coupled soil–leaf-canopy and atmosphere radiative transfer modeling to simulate hyperspectral multi-angular surface reflectance and toa radiance data. *Remote Sensing of Environment*, 109(2):166–182, 2007.

Wout Verhoef, Li Jia, Qing Xiao, and Zhongbo Su. Unified optical-thermal four-stream radiative transfer theory for homogeneous vegetation canopies. *IEEE Transactions on geoscience and remote sensing*, 45(6):1808–1822, 2007.

Eric Vermote, Christopher O Justice, and François-Marie Bréon. Towards a generalized approach for correction of the brdf effect in modis directional reflectances. *IEEE Transactions on Geoscience and Remote Sensing*, 47(3):898–908, 2009.

Dongdong Wang, Shunlin Liang, Yi Zhang, Xueyuan Gao, Meredith GL Brown, and Aolin Jia. A new set of MODIS land products (MCD18): Downward Shortwave Radiation and Photosynthetically Active Radiation. *Remote Sensing*, 12(1):168, 2020a.

Sherrie Wang, Stefania Di Tommaso, Jillian M Deines, and David B Lobell. Mapping twenty years of corn and soybean across the US Midwest using the Landsat archive. *Scientific Data*, 7(1):1–14, 2020b.

Zhuosen Wang, Crystal B Schaaf, Qingsong Sun, Yanmin Shuai, and Miguel O Román. Capturing rapid land surface dynamics with collection V006 MODIS BRDF/NBAR/Albedo (MCD43) products. *Remote Sensing of Environment*, 207: 50–64, 2018.

James Watson, Andrew J Challinor, Thomas E Fricker, and Christopher AT Ferro. Comparing the effects of calibration and climate errors on a statistical crop model and a process-based crop model. *Climatic Change*, 132:93–109, 2015.

Marie Weiss, Frédéric Jacob, and Grgory Duveiller. Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment*, 236:111402, 2020.

J-L Widlowski, Malcolm Taberner, Bernard Pinty, Véronique Bruniquel-Pinel, Mathias Disney, Richard Fernandes, J-P Gastellu-Etchegorry, Nadine Gobron, Andres Kuusk, Thomas Lavergne, et al. Third radiation transfer model intercomparison (rami) exercise: Documenting progress in canopy reflectance models. *Journal of Geophysical Research: Atmospheres*, 112(D9), 2007.

Jean-Luc Widlowski, Corrado Mio, Mathias Disney, Jennifer Adams, Ioannis Andredakis, Clement Atzberger, James Brennan, Lorenzo Busetto, Michaël Chelle, Guido Ceccherini, et al. The fourth phase of the radiative transfer model intercomparison (rami) exercise: Actual canopy scenarios and conformity testing. *Remote Sensing of Environment*, 169:418–437, 2015.

Martin Wild. Enlightening global dimming and brightening. *Bulletin of the American Meteorological Society*, 93(1):27–37, 2012.

Martin Wild, Doris Folini, Christoph Schär, Norman Loeb, Ellsworth G Dutton, and Gert König-Langlo. The global energy balance from a surface perspective. *Climate dynamics*, 40(11-12):3107–3134, 2013.

WMO. WMO Observing Systems Capability Analysis and Review tool (OSCAR). URL `https://www.wmo-sat.info/oscar/`.

CW Wood, DW Reeves, and DG Himelrick. Relationships between chlorophyll meter readings and leaf chlorophyll concentration, N status, and crop yield: a review. In *Proceedings of the agronomy society of New Zealand*, volume 23, pages 1–9, 1993.

George Xian, Collin Homer, and Joyce Fry. Updating the 2001 National Land Cover Database land cover classification to 2006 by using Landsat imagery change detection methods. *Remote Sensing of Environment*, 113(6):1133–1147, 2009.

Qinchuan Xin, Peng Gong, Andrew E Suyker, and Yali Si. Effects of the partitioning of diffuse and direct solar radiation on satellite-based modeling of crop gross primary production. *International Journal of Applied Earth Observation and Geoinformation*, 50:51–63, 2016.

Xiaoxiong Xiong and James J Butler. Modis and viirs calibration history and future outlook. *Remote Sensing*, 12(16):2523, 2020.

J Xu, C Li, H Shi, Q He, and L Pan. Analysis on the impact of aerosol optical depth on surface solar radiation in the shanghai megacity, China. *Atmospheric Chemistry and Physics*, 11(7):3281–3289, 2011.

Xiaojun Xu, Huaqiang Du, Guomo Zhou, Fangjie Mao, Pingheng Li, Weiliang Fan, and Dien Zhu. A method for daily global solar radiation estimation from two instantaneous values using MODIS atmospheric products. *Energy*, 111:117–125, 2016.

Qiangqiang Yuan, Huanfeng Shen, Tongwen Li, Zhiwei Li, Shuwen Li, Yun Jiang, Hongzhang Xu, Weiwei Tan, Qianqian Yang, Jiwen Wang, et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, 241:111716, 2020.

Wenping Yuan, Yang Chen, Jiangzhou Xia, Wenjie Dong, Vincenzo Magliulo, Eddy Moors, Jørgen Eivind Olesen, and Haicheng Zhang. Estimating crop yield using a satellite-based light use efficiency model. *Ecological indicators*, 60:702–709, 2016.

Xiaotong Zhang, Shunlin Liang, Gongqi Zhou, Haoran Wu, and Xiang Zhao. Generating Global LAnd Surface Satellite incident shortwave radiation and photosynthetically active radiation products from multiple satellite data. *Remote Sensing of Environment*, 152:318–332, 2014.

Xiaotong Zhang, Shunlin Liang, Martin Wild, and Bo Jiang. Analysis of surface incident shortwave radiation from four satellite products. *Remote Sensing of Environment*, 165:186–202, 2015.

Yi Zhang, Tao He, Shunlin Liang, Dongdong Wang, and Yunyue Yu. Estimation of all-sky instantaneous surface incident shortwave radiation from Moderate Resolution Imaging Spectroradiometer data using optimization method. *Remote Sensing of Environment*, 209:468–479, 2018.

Xiang Zhao, Shunlin Liang, Suhong Liu, Wenping Yuan, Zhiqiang Xiao, Qiang Liu, Jie Cheng, Xiaotong Zhang, Hairong Tang, Xin Zhang, et al. The Global Land Surface Satellite (GLASS) remote sensing data processing system and products. *Remote Sensing*, 5(5):2436–2450, 2013.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2): 301–320, 2005.